ANALÍTICA DE TEXTO Y PROCESAMIENTO DE LENGUAJE NATURAL APLICADO A NOTAS DE ENFERMERÍA EN ESPAÑOL

ANDRÉS USECHE GOMEZ

FACULTADA DE INGENIERÍA

MAESTRÍA EN ANALÍTICA APLICADA

2022

PÁGINA DE ACEPTACIÓN

Dra. Elena Valentina Gutiérrez
Jurado 1

Luis Gabriel Moreno Sandoval
Jurado 2

Jenny Constanza Robayo Gómez
Jurado 3

William Javier Guerrero Rueda
Tutor

Chía, 5 de agosto del 2022

DEDICATORIA

A mi esposa e hijos, mi motor para seguir adelante

A los profesionales de enfermería, para dignificar su inmensa labor.

AGRADECIMIENTOS

Clínica Universidad de La Sabana.

A los profesores William Guerrero y Luis Gabriel Moreno Sandoval

TABLA DE CONTENIDO

TABLAS

ILUSTRACIONES

PALABRAS CLAVE

Procesamiento de Lenguaje Natural (NLP), Minería de Texto (TDM), Aprendizaje de Máquina (ML), Notas de Enfermería (NN), n-gramas, Lematización, Profesionales de Enfermería (PDE), Clínica de la Universidad de la Sabana (CUS), Datos crudos (Raw Data), Datos Pre-limpios, Datos Limpios, Extracción,-Transformación y Carga (ETL), Palabras Clave, Modelamiento de Tópicos, Asignación latente de Dirichlet (LDA), Frecuencia de Término - Frecuencia Inversa de Documento (TF-IDF), Corpus Estándar De Oro (GSC), Perfilamiento de Autor (AP), CRISP-DM (Proceso estándar interindustrial para la minería de datos), Okapi BM25

Natural Language Processing (NLP),Text Data Mining (TDM), Machine Learning (ML), Nursing Notes (NN), n-grams, lemmatization, Nurse (PDE), Clinic of Sabana's University (CUS), Raw Data, Pre-Clean Data, Clean Data, Extract-Transform-Load (ETL), keywords, Topic Modeling, Latent Dirichlet Allocation (LDA), Term Frequency - Inverse Document Frequency (TF-IDF), Gold Standard Corpus (GSC), Author Profiling (AP), CRISP-DM (Cross Industry Standard Process for Data Mining), Okapi BM25

RESUMEN GRÁFICO

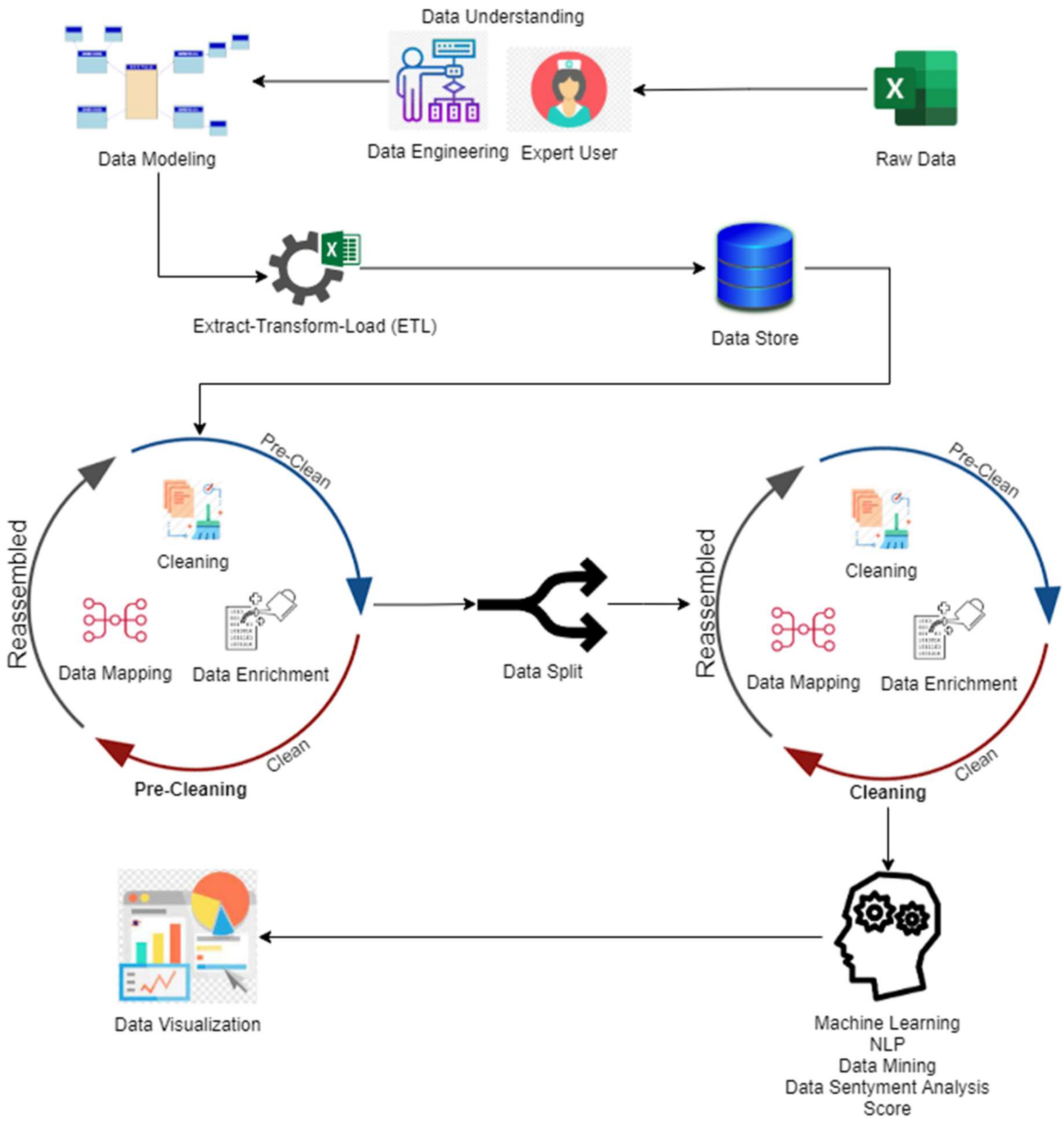


Ilustración 1 - Proceso de Transformación de datos

TÍTULO

Analítica de Texto y Procesamiento de Lenguaje natural aplicado a notas de enfermería en español

RESUMEN

Esta investigación, de tipo exploratoria, tomó una muestra de las notas de enfermería asignada por la Clínica Universidad de La Sabana, se cargaron a una base de datos a través de procesos de extracción, transformación y carga, el texto fue limpiado, corregido y transformado mediante técnicas de limpieza. Con algoritmos de NLP se descubrieron patrones en el estilo de redacción de los enfermeros, se cuantificaron las observaciones registradas con el algoritmo de frecuencia de términos y se detectaron tópicos inmersos en el conjunto de anotaciones que, junto a las categorías previamente revisadas con los profesionales, fueron usadas para agrupar las notas en función de su contenido para calcular una segunda cuantificación usando palabras clave.

Este trabajo está dividido en tres secciones, la primera (Limpieza) ilustra el pre y procesamiento de los datos, la segunda (analítica) usa la información limpia para calcular el puntaje de las notas, el perfilamiento de los enfermeros y categorización de las palabras usadas en los registros, la tercera y última sección (visualización) toma los resultados de los pasos anteriores y los disponibiliza en una herramienta de visualización para exponer la nueva información generada a fin de facilitar el análisis y toma de decisiones

INTRODUCCIÓN

En las entidades prestadoras de salud, la gran cantidad de información registrada diariamente por los Profesionales de Enfermería (PDE) dificulta el proceso de análisis por un humano. Entonces, existe el reto no resuelto de utilizar herramientas de inteligencia artificial para analizar y detectar patrones en el texto, que permita a los tomadores de decisiones, considerar mejores juicios respecto al cuidado de los pacientes y de quienes los atienden. Para el caso de la Clínica Universidad de la Sabana (CUS) existen registros elaborados por los PDE sobre las observaciones realizadas al paciente, teniendo en cuenta su estado físico, emocional, neurológico y hemodinámico, así como los cambios observados, problemas presentados, cuidados brindados, procedimientos realizados, medicamentos suministrados y efectividad de la atención.

Estas notas se registran cronológicamente y son usadas para consultas, material de apoyo a los trabajos de cuidado del paciente y en calidad de documentos científico legal en tribunales de ética [1]. Sin embargo, no son lo suficientemente usadas masivamente, dado que requieren tiempo y recurso humano para su análisis, así mismo, el volumen de información es tan grande, que es inviable hacerlo de forma manual, muchas de ellas no mantienen un formato homogéneo y contienen errores ortográficos por ser un comentario libre diligenciado por humanos. Tecnologías como minería de texto y procesamiento de lenguaje natural pueden reducir la duración requerida para el tratamiento de extensos textos, facilitan encontrar recursos clave con mayor rapidez y eficiencia y propician la generación de nuevo conocimiento a modo de patrones, comportamientos y tendencias.

Este trabajo usó, bajo aprobación del comité de ética en investigación académica del 6 de julio del 2021 de la CUS, muestras retrospectivas de las "Notas de Ingreso" del 9 de mayo al 15 de octubre de 2019 y "Notas de evolución" del 20 de enero al 20 de agosto de 2019 de pacientes adultos en hospitalización sin nombre de paciente ni del personal que los atendió.

Utilizando técnicas de preprocesamiento de texto como limpieza, revisión ortográfica, enriquecimiento de datos y procesamiento de lenguaje natural (NLP), se encontraron interacciones ocultas entre Enfermero-Paciente (E-P) y se identificaron patrones que no eran evidentes en la nota original.

Esta nueva información puede ser empleada por los PDE y áreas administrativas de la CUS para validar la de los registros ingresados, posibilita distinguir aquellos PDE que tienen una baja puntuación acumulada con el contenido implícito en la redacción, ya que cualquier escrito sin importar el idioma soporta ser cuantificado en función del léxico y vocabulario empleado sobre el conjunto total de palabras, también permite correlacionar la cantidad de interacciones reales del E-P y su influencia en la calidad de lo redactado, Así mismo, la cuantificación del texto tolera ser usado para identificar duplicidad de las NN de diferentes pacientes por un mismo enfermero, dado el puntaje que las notas generan.

ESTADO DEL ARTE

En la búsqueda relacionada con la aplicación de minería de texto (TM), aprendizaje de máquina (ML) y NLP para el análisis de las NN, se usó motores de búsqueda especializados como SCIVAL y SCOPUS y se buscaron palabras clave como Nursing Notes, EHR, Nurse Care, Text Classification, NLP, ML, n-grams, entre otras. Como resultado se encontraron varios artículos en Asia, Europa y Estados Unidos, muy pocos de Centro y Suramérica. Estos estudios están enfocados en la prevención de caídas de los pacientes hospitalizados, cálculo de mortandad y estudio a nivel semántico de los registros y su calidad para fines de auditoría.

Como se puede ver en la tabla 1 y la tabla 34 de anexos, los estudios 1, 3 y 5 tienen en común el uso de Minería de Texto (TM) sobre las notas de enfermería, pero solo uno agrupó los datos en tópicos. Las investigaciones 2, 4 y 8 usaron técnicas más avanzadas proporcionadas por el Aprendizaje de Máquina, Inteligencia Artificial y Redes Neuronales para entender el sentido de las palabras usadas por los enfermeros y la transferencia de conocimiento. Los artículos 6 y 7 aplicaron exploración estadística para estimar los tiempos de diligenciamiento de notas y para pronosticar el deceso de un paciente. El artículo 9 ingeniosamente aprovechó información no estructurada como radiografías, exámenes médicos, resultados de laboratorio y demás documentos para complementar el análisis de las anotaciones y el documento 10 nos resalta las consideraciones éticas al realizar este tipo de trabajos.

Al final, de los buenos resultados obtenidos en las investigaciones, se optó por usar Minería de Texto en su versión más ampliada de Analítica de Texto para analizar las notas de enfermería junto a NLP, también se consideró importante incluir tópicos para entender cómo se agrupa el vocabulario usado por los enfermeros. Algo que no tenían los anteriores trabajos e incluidos en este son el estudio en idioma español, el cuantificar el valor de una nota en función de las palabras usadas y de KeyWords inmersas en los registros mediante algoritmos de recuperación de información (IR) e implementar una herramienta de visualización para facilitar el análisis y toma de decisiones.

Tabla 1 - Búsqueda de referencias y tecnologías/aspectos trabajados

N°	PAPER	Text Mining Text Analytics	Electronic Health Records	Topic Modeling	Machine Learning	Artificial Intelligence	N-grams	Lexicon	LIWC	Human Process	Statistic Analysis	Neuronal Networks	Unstructured Data	Spanish Language	Score	Visualization	Quality	Natural Language Processing	Ethics
1	Applicaton of Text Ming to Nursing Texts	x	x	x															
2	Testing the Use of Natural Language Processing Software and Content Analysis to Analyze Nursing Hand-off Text Data				x				x										
3	Using a Text Mining approach to explore the recording quality of a nurse record system	x	x														x		
4	What Can We Learn about Fall Risk Factors from EHR Nursing Notes? A text Mining Study				x		x	x											
5	A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data	x	x															x	
6	Using nursing notes to improve clinical outcome prediction in intensive care patients: A retrospective cohort study									x									
7	The influence of integrated electronic medical records and computerized nursing notes on nurses' time spent in documentation										x								
8	Applying artificial intelligence technology to support decisionmaking in nursing: A case study in Taiwan					x						x							
9	Challenges and opportunities beyond structured data in analysis of electronic health records										x		x						
10	Theoretical Considerations of Ethics in Text Mining of Nursing Documents										x								x
11	This approach	x	x	x			x			x				x	x	x	x	x	x

PREGUNTA DE INVESTIGACIÓN PRINCIPAL

¿Es posible generar nuevo conocimiento a partir de las notas de enfermería en español de la Clínica Universidad de La Sabana y de cualquier otra, mediante técnicas de procesamiento de datos y analítica de texto y la construcción de indicadores de mejora continua de quien las elaboró y del contenido?

PREGUNTA DE INVESTIGACIÓN DERIVADAS

- ¿Permitirá la analítica de texto generar nueva información que ayude a los profesionales de enfermería en su trabajo?
- ¿Podrá la categorización de las notas de enfermería ser un indicador para evaluar su contenido?
- ¿Será el puntaje el mecanismo ideal que permitirá ranquear la calidad de las notas de enfermería y habilidades de redacción del enfermero que las realizó?

OBJETIVOS

General

Demostrar la posibilidad de generar nuevo conocimiento a partir de las notas de enfermería en español de la Clínica Universidad de La Sabana y de cualquier otra, mediante técnicas de procesamiento de datos y analítica de texto y su potencial uso para la construcción de indicadores que permitan la mejora continua de quien las elaboró y del contenido que deberían tener.

Específicos

- Estructurar los procesos necesarios que permitan hacer la limpieza de las NN, para que posteriormente puedan ser usadas en la creación de indicadores para la toma de decisiones.
- Generar puntajes sobre las notas de enfermería en función de su contenido para evaluar su calidad.
- Proponer la arquitectura técnica y buenas prácticas para disminuir la curva de aprendizaje y tiempos requeridos en la implementación de proyectos de este tipo

MARCO CONCEPTUAL

Para el desarrollo de la investigación es necesario tener en cuenta varios niveles de conocimiento, según se avanza en cada una de las tres secciones. Para la primera sección es necesario conocer sobre ETLs, manipulación de datos, expresiones regulares, SQL, manipulación de grandes volúmenes de información y técnicas de limpieza. En la segunda, con los datos procesados, se requiere conocer acerca de algoritmos de recuperación de información, como TD-IDF, Okapi BM25, Author Profiling y Python. La tercera requiere técnicas de visualización de datos y de contar historias a partir de los datos, aparte de contar con el conocimiento para realizar tableros, reportes y KPI’s. Los relevantes a continuación:

Acortamiento (Stemming)
Este proceso heurístico está asociado a la lingüística, que consiste en recortar las palabras según las reglas gramaticales de cada idioma a su base común (Ej. Comienza=comienz, comenzarán=comenz, clases=clas, corrieron=corr). A diferencia de la lematización, el acortamiento es computacionalmente más rápido que la lematización, aunque no es tan preciso gramaticalmente.[24]

Expresiones regulares (Regular Expressions)
Las expresiones regulares son una secuencia de caracteres que conforman un patrón de búsqueda, son de gran utilidad para encontrar ese patrón de combinación dentro de una cadena de texto

Extracción, Transformación y Carga (ETL)
Proceso mediante el cual se obtienen datos de una fuente (Extracción), se procesan (Transformación) en un formato que pueda ser consultado y se persisten (Cargan) en una base de datos relacional, bodega de datos u otro sistema para su posterior uso.

Lematización (lemmatization)
La lematización es un proceso lingüístico que consiste en hallar el lema de una palabra, donde los nombres y adjetivos son convertidos a su masculino y singular y los verbos en infinitivo, en su forma canónica (Ej. Comienza=comenzar, comenzarán=comenzar, clases=clase, corrieron=correr). Es un proceso que consume muchos recursos, en especial tiempo y dado que es probabilística, puede generar algunos problemas inesperados. El objetivo de la lematización y el acortamiento es reducir la cantidad de tokens únicos, lo cual es apreciado en análisis de contenido.[23]

N-gramas (n-grams)
Un n-grama es una subsecuencia de *n* elementos de una secuencia dada (o grupo de tokens), los elementos pueden ser números, símbolos, fonemas, sílabas, letras y palabras. La letra *n* representa el número de secuencias a tomar, los hay de 1-grama o unigramas (Ej. Paciente, enfermero, cc, 10, ml), 2-gramas (sin riesgo, con riesgo, cuidado médico) o bigramas, 3-gramas o trigramas (sin riesgo de caída, asistencia médica complementaria), y así, sucesivamente. Los n-gramas son usados en análisis de texto para convertir un texto de formato no estructurado en estructurado, también , a partir de los n-gramas, se pueden construir una bolsa única de n-gramas, la cual cuenta la cantidad de veces que aparecen. Una vez generados, se pueden utilizar con los algoritmos de ML para crea modelos predictivos.[22]

Palabras y Frases Reservadas (Stop Words and Stop Phrases)
Las palabras y frases reservadas son tokens que carecen de sentido para nuestro análisis textual, se pueden considerar como palabras irrelevantes (artículos, pronombres o palabras customizadas), también algunos signos de puntuación y símbolos se pueden considerar stopwords (Según el enfoque de investigación que se quiera realizar). [23]

Ejemplo de pronombres
Cualesquiera, Cualquiera, Demasiadas, Demasiados, Demasiada, Demasiado, Vosotras, Aquellas, Aquellos, Ningunas, Ningunos, Nosotras, Nosotros, Nuestras, Nuestros, Vosotros, Vuestras, Vuestros, Alguien, Algunas, Algunos, Aquella, Aquello, Conmigo, Consigo, Contigo, Escasas, Escasos, Ninguna, Ninguno, Nuestra, Le, Lo, Me, Mi, Os, Se, Si, Te, Ti, Tú, Yo.

Ejemplo de artículos
Como Se Mencionó Anteriormente, Con la Condición De Que, En Lo Que Respecta A, Con La Intención De, Y Así Sucesivamente, Con El Objetivo De, De La Misma Manera, Después De Lo Cual, En Comparación Con, A Consecuencia De, Con El Fin De Que, De Tal Manera Que, En Otras Palabras, En Pocas Palabras, Como He Mostrado, En Lo Relativo A, Él Como, Para, Pero, Pues, Si no, Así, Si, Y, O.

Ejemplo de Frases
El catálogo de frases a excluir es un proceso que se debe realizar con los auxiliares de enfermería, sin embargo, se diseñó un catálogo para almacenarlas. (Ver tabla 2)

Tabla 2 - Stop Phrases

Generic_Data_Key	Text_Value	Text_Type	Source	Active	Load_DT	Notes
1	PACIENTE DESPIERTO ALERTA Y TRANQUILO	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
2	INTERVENCION INMEDIATA	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
3	INGRESO HOSPITALIZACION POR ENFERMERIA	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
4	SE REALIZA PROTOCOLO DE BIENVENIDA	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
5	DIAGNOSTICO DE ENFERMERIA	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
6	FINALMENTE SE DA ESPACIO PARA HACER PREGUNTA Y RE...	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
7	SE ABRE FOLIO PARA REGISTRO DE CONTRASTE PARA TAC	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
8	PLAN DE ENFERMERIA	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
9	INFORMACION DE INGRESO	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
10	SE EDUCA EN CUIDADO Y AUTOCUIDADO	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
11	POLITRAUMATISMO	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
12	VALORACION E IDENTIFICACION DE RIESGOS	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
13	SE ABRE FOLIO PARA MEDIO DE CONTRASTE	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
14	CONTROL DE SIGNOS VITALES	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
15	ASISTENCIA EN NECESIDADES BASICAS	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL
16	SIN APARENTE DEFICIT NEUROLOGICO AFEBRIL	Phrase	Notas_Ingreso	1	2021-07-29 09:11:06.953	NULL

Revisor Ortográfico (Spell Check)
Un corrector ortográfico es una herramienta usada en el análisis de texto con el fin de detectar faltas ortográficas y corregirlas de una manera automática o manual. Esta corrección es importante cuando se detecta que el origen de los datos son registros ingresados libremente por usuarios, sin ningún tipo de filtro. Al aplicar el corrector ortográfico, facilitamos el análisis de contenido de los pasos posteriores. [22]

Tokenización (Token)
La Tokenización es una técnica usada para segmentar un texto en tokens, un token es un conjunto de caracteres que representan el texto o es la unidad de análisis del texto. Un token no es lo mismo que una palabra, la palabra posee un significado por sí misma, mientras que un token es un elemento abstracto. Normalmente, existe un carácter que separa un token de otro, regularmente es el símbolo de espacio. Tokenizar permite identificar la frecuencia de apariciones de un token, eliminar

tokens que no aportan significado (stop words) y realizar correcciones ortográficas.[23]

Análisis cualitativo y cuantitativo de contenido
Las NN son texto libre no estructurado en lenguaje español de tipo cualitativo elaborado por los PDE, dentro de esta etapa se examina cantidad de frases, cantidad de frases únicas en las NN, total de palabras, total de palabras únicas (vocabulario), estilo de redacción (escribe en solo mayúsculas, en solo minúsculas, usa mayúsculas y minúsculas al tiempo, que tanto usa caracteres especiales), análisis de contenido, Perfil del autor, Rasgos Léxicos, Diversidad Lexical.

Concurrencias (Co-occurrence)
Las reglas de Coocurrencias permiten descubrir y agrupar conceptos fuertemente relacionados dentro del conjunto de documentos o registros. Cuando esa Coocurrencia se hace más fuerte dentro del documento, puede denotar una relación subyacente para la creación de categorías. [25]

Corpus Estándar De Oro (GSC)
El GSC en el contexto de NLP es la colección manual de anotaciones de texto evaluadas por múltiples expertos, donde previamente los datos son fueron evaluados individualmente y al final pasaron por un consenso para establecer qué palabras deberían estar o no dentro del diccionario final, también son el filtro que permite detectar ambigüedades o sociolectos que los algoritmos no pueden reconocer.[30]

Modelado de Tópicos (Topic Modeling)
Es una técnica no supervisada de NLP es capaz de detectar relaciones semánticas en grandes volúmenes de información, estas relaciones, llamadas tópicos, estos son un conjunto de palabras que suelen aparecer juntas en los mismos contextos y nos permiten observar relaciones no evidentes. [27].

El algoritmo usado en el modelo Latent Dirichlet Allocation (LDA) [28], propone los diferentes tópicos que componen las notas y cada cuanto está presente cada tópico en los documentos. El modelamiento de tópicos más las categorías definidas por los PDE son el medio principal usado por esta investigación para resaltar la metadata en las NN.

Con las notas lematizadas, estas son cargadas a Python, donde son procesadas y agrupadas por el número óptimo de tópicos calculados mediante la función compute coherence values [27]. Las Keywords más relevantes de cada tópico son usadas para alimentar la tabla

Okapi BM25
Es una variación de TF-IDF usada para recuperación de información, en ella se mejora el algoritmo al penalizar el cálculo en virtud de la longitud del texto, entre más largo el documento, más es el castigo. Estos modelos asumen que las apariciones de un término en un documento tienen una naturaleza aleatoria, de tal forma que un documento es visto como una secuencia aleatoria de términos. [29]

Recursos Lingüístico

Un recurso lingüístico es un conjunto de corpus, diccionarios y lista de palabras de un lenguaje en particular, el cual puede ser usado como referencia, por citar un solo ejemplo, el análisis de sentimientos (minería de opiniones) entre otros; para mejorar la fuerza de los léxicos y disminuir su debilidad [13, 14, 15].

Construir recursos lingüísticos es costoso por la cantidad de esfuerzo y tiempo para construirlos. Actualmente, se encuentra gran cantidad de recursos lingüísticos para el idioma inglés, pero para el español y otros idiomas es más difícil encontrarlos. Estos recursos se reducen sin son de tipo Open Source. Una estrategia para aumentar la base de conocimiento es agrupar varios lexicones (fuentes de información con una temática definida), pero para ello se deben ejecutar procesos de consolidación, de limpieza, como eliminar palabras, frases repetidas, y retirar aquellas que tienen menor probabilidad de uso.

La importancia de estos recursos lingüísticos obedece a que, con ellos, se puede mejorar la calidad y predictibilidad de los modelos (Precisión and recall), también pueden ser usados como diccionarios o valores de referencia [16]. Para las notas de enfermería y terminología médica, no es fácil encontrar recursos lingüísticos en español, sin embargo, recursos como UMLS y SNOMEDCT-SP son fuentes importantes que pueden ser usados para dicho fin [3].

TF-IDF

La frecuencia de término - frecuencia inversa de documento es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. Esta medida se utiliza a menudo como un factor de ponderación en la recuperación de información y la minería de texto. El valor tf-idf aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia de la palabra en la colección de documentos, lo que permite manejar el hecho de que algunas palabras son generalmente más comunes que otras. [29]. En el análisis tradicional de palabras clave, se mide qué tan frecuente es una palabra, pero por lo general, muchas de ellas como los stopwords se usan con más frecuencia, pero no aportan para comprender los temas relevantes. La tf-idf permite dar contexto más amplio a la palabra, según al texto completo y al número de documentos. Ver fórmula matemática en [29].

MARCO METODOLÓGICO

Este proyecto usa la metodología CRISP-DM [33] , este es un estándar de trabajo para proyectos de minería de datos, la cual comprende una serie de fases (Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación y Despliegue) que ejecutadas correctamente aumentan la probabilidad de éxito del proyecto.

Para la fase de comprensión del negocio se realizó encuesta con los profesionales de enfermería acerca del uso actual de las NN, los problemas más habituales y el uso adicional al actual que podrían tener, también se realizó reuniones con los PDE para comprender más de cerca el proceso para realizarlas.

Para la fase de comprensión de datos y por el volumen de registros, fue necesario cargar los datos crudos desde archivos de Excel hacia una base de datos mediante ETLs, en ella fue se generó estadísticas como longitud mínima, media y máxima de las NN, cantidad de palabras usadas, cantidad de registros cargadas, valores nulos presentes por atributo, entre otros.

Para la fase de preparación de los datos, fue necesario dividir las notas en tokens para entender la frecuencia que tienen en el conjunto de datos, se encontraron problemas de ortografía, uso sin estándar de siglas y acrónimos, y uso de lenguaje médico, 80% del tiempo invertido en el proyecto se ejecutó en esta etapa.

En la etapa de modelado y con los datos limpios se usaron algoritmos de NLP para analizar lingüísticamente el vocabulario usado, con Topic Modeling se analizó cómo se distribuyen los tokens en tópicos y con estos, se cuantificaron el valor de cada nota en función de palabras clave. También se diseñaron tableros y reportes para interactuar con la información generada.

Este proyecto no incluye la etapa de Evaluación y despliegue, ya que están para una investigación futura.

Historias de Usuarios

Las historias de usuarios hacen parte de metodologías ágiles para el desarrollo de proyectos y tiene como finalidad entender la situación actual del diligenciamiento de las NN, que funciona y que no, así mismo, da una perspectiva en el ámbito de usuario final de lo que puede mejorar. Para lo anterior se realizó encuesta a los profesionales de enfermería, las preguntas fueron previamente validadas por sesiones uno a uno con la jefe de enfermería, luego se hicieron encuestas unitarias con tutor de la tesis para validar percepción y finalmente se habilitó el cuestionario vía formulario electrónico al líder de enfermeros, quien lo socializo con los PDE. Las interrogantes planteadas son las siguientes:

¿Para qué usa las notas de enfermería actualmente?, ¿Qué problemas ha tenido al leer las notas de enfermería? (Ortografía, siglas, Convenciones, Normas), que otras?, Si se hace un proceso exhaustivo de limpieza en las notas de enfermería (Corrección Ortográfica y de abreviaciones), ¿cree que podría usar con más frecuencia las notas de enfermería?, ¿Cree que las notas de enfermería limpias y resumidas podrán ayudarle a agilizar el proceso de entrega de turno?, ¿Qué información mínima debería tener la nota de enfermería? (Estado de conciencia, riesgo de caída, acompañante, dispositivos), ¿alguna otra?, ¿Le parecería útil tener un puntaje de las notas de enfermería, con el fin de calificar las notas en función de palabras clave que deberían tener la nota?, ¿Considera que las notas de enfermería podrían tener otro uso al del actual? ¿Cuál sería?, ¿Podría la nota de enfermería limpia y resumida mejorar las estrategias de atención a los pacientes, como se podría lograr?, ¿Cuáles son las consideraciones técnicas que debe tener cualquier herramienta tecnológica para que sea fácilmente implementada bajo la actual plataforma tecnológica con la que cuenta la Clínica de la Universidad de la Sabana?

La encuesta fue analizada mediante técnicas de NLP y Topic Modeling (Ver ilustración 2) y mediante inspección manual por un humano (Ver ilustración 3) con lo cual se identificó los temas más relevantes revelados por los encuestados.

Análisis de la encuesta mediante NLP y Topic Modeling
Para el análisis de la encuesta con NLP y TM se usó agrupación de tokens por trigramas, bigramas y unigramas (Ver ilustración 2). Los círculos revelan las agrupaciones que están fuertemente correlacionadas e inmersas en la encuesta.

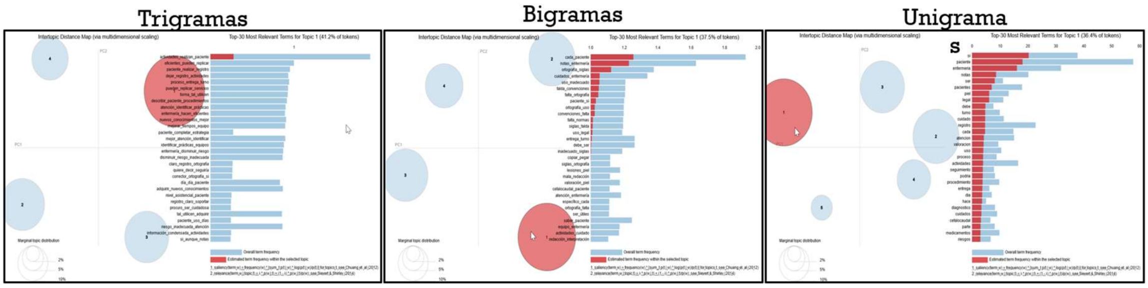


Ilustración 2 - Topic modeling por trigramas, bigramas y unigramas aplicado a la encuesta realizada a los PDE

Análisis de la encuesta mediante proceso manual
El análisis manual de la encuesta fue tabulado y se obtuvo las palabras/frases más frecuentes por pregunta, para determinar las variables predominantes. La ilustración 3 muestra aquellas cuyo conteo es mayor a 3.

Respuestas	Pregunta 2	Pregunta 7	Pregunta 4	Pregunta 6	Pregunta 3	Pregunta 8	Pregunta 1	Pregunta 5	Grand Total
Procedimientos, actividades y tratamientos (realizados y por realizar) al paciente		2	2	3	8		25	8	48
Si parece útil el score				35					35
Agilizar la entrega de turno			30		2				32
Estado clínico actual del paciente				3	3	4	14	7	31
No se considera otros usos de las NN		24							24
La limpieza y resumen de la nota SI agregan valor						20			20
Corrección de errores de ortografía	12			1	5				18
La limpieza y resumen de la nota NO agregan valor						18			18
Si considera otros usos a las NN		15							15
Valoración de piel del paciente								14	14
Unificar siglas y abreviaciones en las NN	7				5	1			13
Administración de medicamentos al paciente (Buena o Mala)						2	7	1	10
Unificación en la redacción de las NN	8			1		1			10
Como soporte para investigaciones		9							9
Cuidados de enfermería y cambios del paciente en el turno			2			1	2	4	9
No agilizan la entrega de turno			9						9
Valoración Cefalocaudal del paciente			2				1	6	9
Dispositivos usados en el paciente							3	5	8
Evolución del paciente (Bueno o malo)		1				1	4	2	8
Reducción de tiempo en el sistema al diligenciar la NN				1	3	3			7
Riesgo del paciente			1			1	2	3	7
Respaldo legal					1	1	4		6
Cuidados del paciente							5		5
Definir lenguaje y normas para diligenciar las NN	2				1	2			5
Disminución, revisión y facturación de glosas		2		1		2			5
Notas que no corresponden al paciente	5								5
Plan de cuidado personalizado del paciente según su estado			1			2		2	5
Registro de NN cronológicamente					1		4		5
Signos vitales del paciente					1		2	2	5
No hay utilidad en el Score				4					4
Plan de enfermería personalizado según diagnostico del paciente			1		1			2	4
Posición del paciente					2		1	1	4
Actividades de enfermería							3		3
Alergias del paciente						1		2	3
Antecedentes del paciente			1			1		1	3
Estado de conciencia del paciente								3	3
Insumos usados en el paciente							3		3
Mas tiempo de atención al paciente						3			3
Ninguna	3								3
No habría cambio					3				3
Riesgo de Caída del paciente			1					2	3
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Grand Total	52	56	57	58	62	85	93	101	564

Ilustración 3 - Tabulación manual de la encuesta realizada a los PDE sobre el uso de las NN

NOTA: La pregunta 9 fue excluida por ser solo para el área técnica.

El análisis conjunto entre NLP y revisión manual resaltó las variables más predominantes en la encuesta. El resultado revela enfoques en dos aspectos, hacia la nota y la calidad de contenido que deben tener y la otra en cuanto a la información del paciente que debería tener. (Ver tabla 3).

Tabla 3 - Variables más relevantes detectadas por el análisis de la encuesta

Sobre las notas de enfermería	**Sobre el paciente**
Les falta calidad	Deben contener sus antecedentes
Tienen errores	Deben contener sus riesgo de caída
Son usadas como respaldo legal	Deben contener su Estado de Conciencia
Tienen problemas de ortografía	Deben contener Dispositivos usados en el
Tienen problemas de Redacción	Deben contener su Estatus actual
No tienen un estándar en la redacción	Deben contener su Evolución
No hay estándar en el uso de siglas, acrónimos, abreviaciones y convenciones	Deben contener los Insumos utilizados en el
Deben contener el registro de actividades realizadas	Deben contener los Medicamentos suministrados
Deben contener los Procedimientos Realizados o a realizar	Deben contener su Valoración de Piel
Deben contener los Cuidados de Enfermería realizados	Deben contener los tipos de riesgos que pueden tener
Deben contener información útil para las entregas de turno	Deben contener sus Signos Vitales
Seria ideal poder cuantificar los cuidados de enfermería y poder generar nuevos conocimientos a partir de ellos	Deben contener su Valoración

Mediante la metodología de prospectiva y la herramienta MICMAC [31], se analizaron las variables para detectar las más influyentes. (Ver ilustración 4)

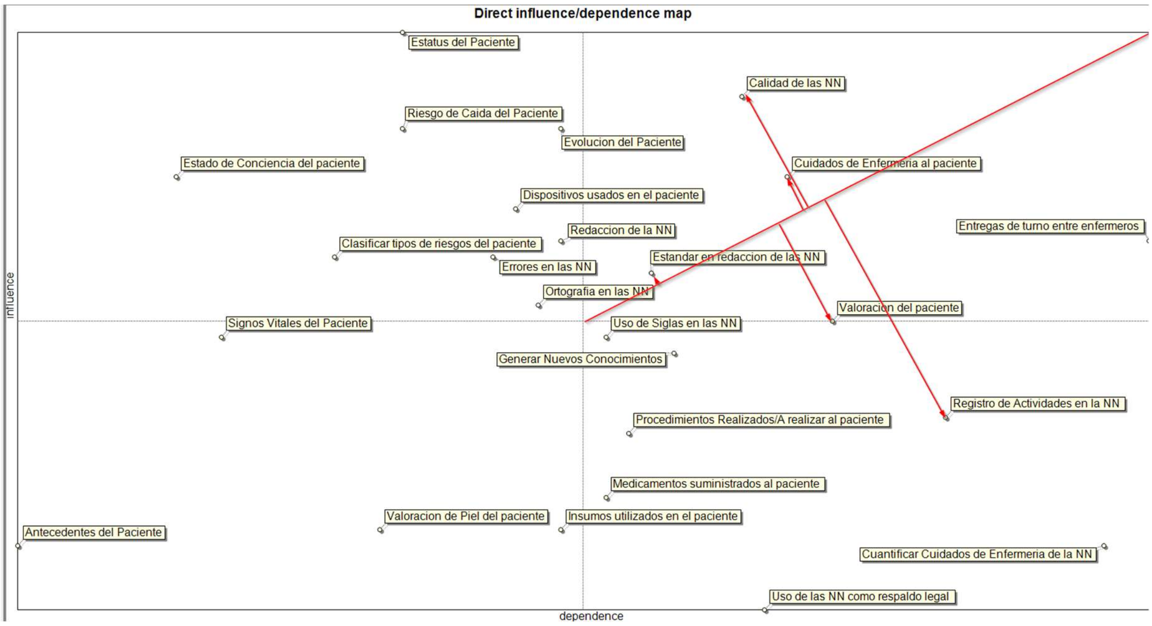


Ilustración 4 - Variables influyentes generadas por la herramienta MICMAC

Los cuidados de enfermería y las entregas de turno son las variables más representativas y que influencian directamente otras variables, sin embargo, son temas que se deben tratar en gestión del cambio. Desde el punto de vista de analítica de datos, la calidad, la ortografía, los errores, la redacción y el estándar en las notas de enfermería son puntos clave que se pueden mejorar con algoritmos de NLP. El entendimiento de las necesidades de los PDE sobre las NN cimienta el camino de cómo abordar el proyecto.

Solicitud de datos

Dada la naturaleza confidencial de las notas de enfermería de la Clínica de la Sabana, la junta del comité de ética aprueba el 6 de julio del 2021 el acceso a los datos, con las recomendaciones y tratamientos de datos personales pertinentes.

Recolección de datos

Los datos usados pertenecen a un muestreo por conveniencia de los sistemas informáticos de la CUS, previa autorización del comité de ética. Este proyecto no conocía los datos ni sugirió ninguna fecha en particular ya que son irrelevantes en la investigación, la información proporcionada es la siguiente:

- Histórico de los signos vitales de los pacientes desde 12 de enero de 2019 al 18 de agosto de 2020 y la hora de captura (presión sistólica, presión diastólica, presión venosa central, temperatura, frecuencia respiratoria, frecuencia cardiaca, pulso, Glasgow, peso, talla, estado de hidratación, personal que toma signos, codificada y anonimizada, examen físico de piel, saturación de O2, examen neurológico, examen respiratorio, escala de dolor). Sin nombre de paciente ni de personal que lo realiza.
- Notas de ingreso de enfermería del 9 de mayo de 2019 al 15 de octubre de 2019 en hospitalización adultos. Sin nombre de paciente ni de personal que lo realiza.
- Notas a la evolución de enfermería de los pacientes del 20 de enero de 2019 al 20 de agosto de 2019 para pacientes. Sin nombre de paciente ni de personal que lo realiza.
- Histórico desde el 16 de enero de 2019 al 19 de julio de 2020 del detalle de los líquidos de cabecera suministrados a los pacientes, frecuencia, cantidad. Sin nombre de paciente ni de personal que lo administra.
- Histórico desde el 16 de enero de 2019 al 19 de julio de 2020 del detalle de los líquidos suministrados a los pacientes en hospitalización adultos (sin nombre de paciente ni de personal que lo administra). Cantidad aplicada, estado del registro y hora del registro.
- Caracterización demográfica de los pacientes hospitalizados en las fechas de las notas de ingreso (municipio de residencia, barrio, estado civil, nivel educativo, ocupación, departamento de nacimiento, fecha de nacimiento, discapacidad, etnia, grupo poblacional).

Consideraciones Éticas

- Este estudio tendrá en cuenta lo dispuesto en las leyes 266 de 1996, y 911 de 2004 por la cual se reglamenta el ejercicio de enfermería y se instaura su código. Allí se establece que los procesos de investigación deben salvaguardar la dignidad, la integridad y los derechos de los seres humanos, como principio ético fundamental y la resolución número 1995 de 1999 por la cual se establecen normas para el manejo de la Historia Clínica.

- Además, se contemplará lo dispuesto en la Resolución 008430 de 1993, por la cual se establecen las normas científicas, técnicas y administrativas para la investigación en salud (Ministerio de salud, 1993).
- Se garantizará la privacidad de los participantes y se tendrá en cuenta los principios de beneficencia, no maleficencia, autonomía, justicia, veracidad, fidelidad y reciprocidad. Toda la información recolectada será anonimizada y se custodiará exclusivamente en los archivos privados del investigador principal por un periodo de cinco (5) años. Luego de este tiempo, toda la información será destruida y borrada de los archivos privados.
- No se harán publicaciones con información privada o confidencial.
- Además, los derechos patrimoniales de los desarrollos tecnológicos asociados al proyecto son de propiedad de la Facultad de Ingeniería de la Universidad de La Sabana y son intransferibles, según la política de propiedad intelectual de la Universidad.
- Con relación a las publicaciones y participación en eventos científicos y culturales, se harán de manera conjunta y contarán con el aval del director del proyecto. Los autores se determinarán según los aportes al proyecto y a la construcción de este, dentro del marco de las políticas de propiedad intelectual de la Universidad de La Sabana.
- Por su naturaleza, el estudio no tiene un impacto ambiental negativo. Sin embargo, atendiendo la política de Compromiso con el Medio de la Universidad y de la Clínica, sus autores se comprometen con el uso responsable de recursos tales como el papel, la energía y manejo adecuado de desechos, en los cuales además trabajarán aportando a futuro por la naturaleza de la propuesta que se diseña.

Marco Legal

- El estudio fue planificado de modo tal que en que no se sometan a los pacientes a riesgo innecesarios y se rige por los preceptos éticos recomendados por la Asociación Médica Mundial (Helsinki, 1964; Tokio, 1975; Venecia, 1983; Hong Kong, 1989; Sudáfrica 1996; Escocia, 2000; Washington, 2002; Tokio, 2004 y Seúl, 2008).
- El investigador principal y coautores en representación de la Clínica Universidad de la Sabana aseguran que la información relativa al estudio no está disponible a personas no autorizadas para el mismo, así como los datos personales del paciente, se mantendrán bajo confidencialidad de conformidad con la (Resolución 8430/1993, Decreto 2378/2008) , política de seguridad del paciente, y normatividad vigente y a la libre circulación de estos datos.
- El investigador principal y los investigadores secundarios, en representación de la Clínica Universidad de la Sabana, aseguran los derechos y libertades de las personas físicas, en lo que respecta al procesamiento de datos personales, y en particular, a su derecho de la intimidad.
- El investigador principal y los investigadores secundarios proporcionan a la Clínica Universidad de la Sabana y (En caso de desarrollar el estudio en conjunto con otra institución Ej. Universidad, Instituto) solo datos hechos anónimos.

- El investigador principal y los investigadores secundarios podrán utilizar los datos analizados y provistos de acuerdo con los cronogramas establecidos y aprobados del proyecto/estudio/protocolo.
- En todas las publicaciones de trabajo en revistas científicas

Reconocimiento de Texto

La información fue recibida en archivos de Excel, el idioma usado es el español, con algunas pocas palabras en inglés, el dominio de conocimiento es de tema médico

Modelo de Datos

La ilustración 5 muestra el modelo de datos sugerido para almacenar la información cruda y procesada de las NN, así como las tablas auxiliares requeridas en el proceso. En la izquierda se encuentran las entidades donde se carga las notas crudas (staging area), tal y como llegan de la fuente, y en la parte derecha se encuentran las tablas paramétricas que hacen parte del proceso de transformación y enriquecimiento de datos.

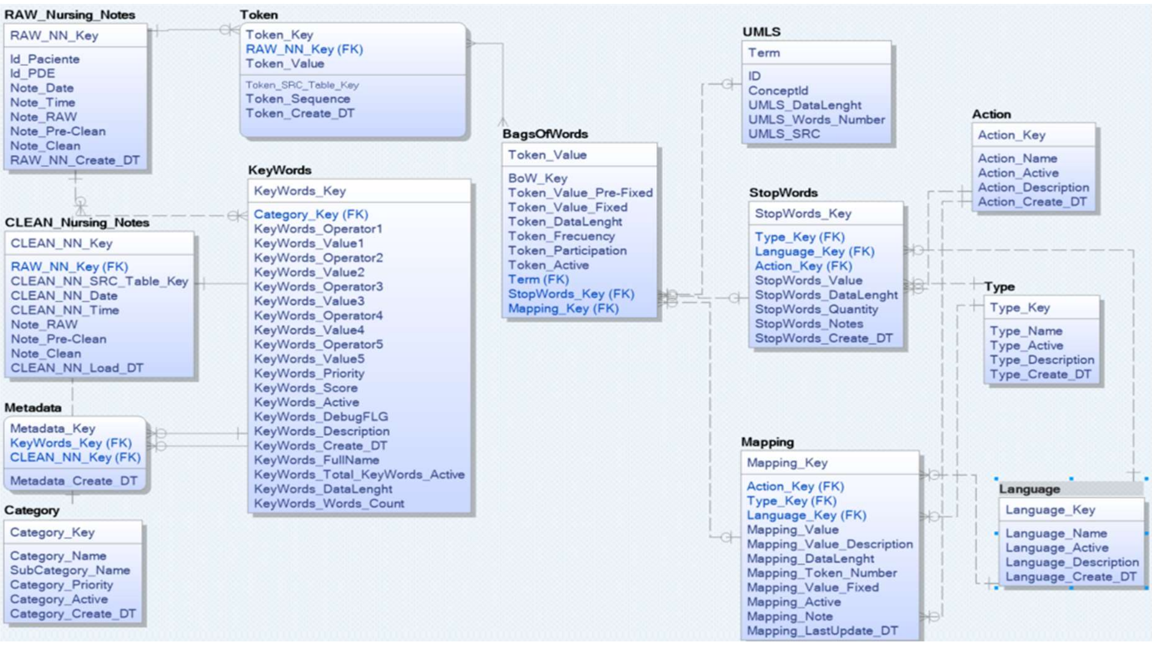


Ilustración 5 - Erwin Data Model

Recepción y Carga de Datos

La información fue recibida en archivos de Excel y mediante un proceso de ETL, fue cargada a la base de datos SQL Server 2019, los archivos recibidos son:

- Notas de Ingreso (información del paciente al momento de llegar a la CUS, esto incluye valoración inicial por parte del PDE)
- Notas de Evolución (Información de la evolución del paciente según los registros diligenciados por los PDE en cada una de sus visitas)
- Caracterización del paciente (Datos demográficos y socioculturales del paciente)

- Signos Vitales (Signos vitales del paciente)
- Líquidos Cabecera (Resumen de los líquidos administrados al paciente)
- Líquidos Detalle (Detalle de los líquidos administrados al paciente)

Sin embargo, las únicas fuentes de información cargadas que contienen notas de enfermería son las Notas de Ingreso y las Notas de Evolución. Adicionalmente, se crearon dos columnas artificiales, Notas_PRE-CLEAN y Notas_CLEAN, las cuales serán usadas para almacenar los cambios de limpieza de datos a partir de los datos originales (Ver Tabla 4).

Tabla 4 - Notas de enfermería originales para ingreso y evolución

Origen	Notas_Ingreso_Key	Upload_DT	IdPaciente	FechaHoraRegistro	UsuarioEvoluciona	Nota_RAW
Notas_Ingreso	523	2022-05-08 13:43:00	Paciente 577	2019-12-26 18:14:32.000	NULL	se abre folio para insumo gastro
Notas_Ingreso	524	2022-05-08 13:43:00	Paciente 577	2019-12-26 18:15:27.000	NULL	se abre folio para solicitar insumo gastro
Notas_Ingreso	525	2022-05-08 13:43:00	Paciente 577	2020-01-06 22:00:07.000	NULL	*** MOMENTO DE CUIDADO - INGRESO** 07/01/2020 INGRESA PACIENTE ...
Notas_Evolucion	672	2022-05-08 13:44:00	Paciente 577	2019-01-12 00:00:00.000	AU555	ADMINISTRACIÓN Y EDUCACIÓN DE MEDICAMENTOS Previa valoración i...
Notas_Evolucion	673	2022-05-08 13:44:00	Paciente 577	2019-01-13 00:00:00.000	AU493	22+00 Se realiza ronda de enfermería, se encuentra paciente en la unidad, bajo ...
Notas_Evolucion	674	2022-05-08 13:44:00	Paciente 577	2019-01-13 00:00:00.000	AU555	VALORACIÓN DEL PACIENTE /SEGUIMIENTO DE ENFERMERÍA PACIENTE ...
Notas_Evolucion	675	2022-05-08 13:44:00	Paciente 577	2019-01-13 00:00:00.000	AU493	se abre folio por error
Notas_Evolucion	676	2022-05-08 13:44:00	Paciente 577	2019-01-13 00:00:00.000	AU555	ADMINISTRACIÓN Y EDUCACIÓN DE MEDICAMENTOS Previa valoración i...
Notas_Evolucion	677	2022-05-08 13:44:00	Paciente 577	2019-01-13 00:00:00.000	AU493	05+00 Se realiza ronda de enfermería, se encuentra paciente en la unidad, bajo ...
Notas_Evolucion	678	2022-05-08 13:44:00	Paciente 577	2019-01-13 00:00:00.000	AU450	07+00 Recibo paciente de 85 años de edad en unidad con barandas elevadas, m...
Notas_Evolucion	679	2022-05-08 13:44:00	Paciente 577	2019-01-13 00:00:00.000	EN289	***ADMINISTRACION DE MEDICAMENTOS*** Previa valoración integral del pa...
Notas_Evolucion	680	2022-05-08 13:44:00	Paciente 577	2019-01-13 00:00:00.000	AU450	08+00 Paciente acepta y tolera via oral 09+00 Se realiza toma y registro de sign...
Notas_Evolucion	585	2022-05-08 13:44:00	Paciente 577	2019-01-09 00:00:00.000	AU450	13+00 Paciente acepta y tolera dieta ordenada 14+30 Se realiza toma y registro...
Notas_Evolucion	586	2022-05-08 13:44:00	Paciente 577	2019-01-09 00:00:00.000	AU450	18+30 Paciente acepta y tolera via oral Entrego paciente de 85 años de edad e...
Notas_Evolucion	587	2022-05-08 13:44:00	Paciente 577	2019-01-09 00:00:00.000	EN204	se abre folio por error
Notas_Evolucion	588	2022-05-08 13:44:00	Paciente 577	2019-01-09 00:00:00.000	EN065	19+00 Recibo paciente de 85 años de edad con diagnóstico: hemorragia de vias...
Notas_Evolucion	589	2022-05-08 13:44:00	Paciente 577	2019-01-09 00:00:00.000	AU555	ADMINISTRACIÓN Y EDUCACIÓN DE MEDICAMENTOS Previa valoración i...
Notas_Evolucion	590	2022-05-08 13:44:00	Paciente 577	2019-01-09 00:00:00.000	AU555	EDUCACIÓN A PACIENTE Una vez identificadas las necesidades educativas d...
Notas_Evolucion	603	2022-05-08 13:44:00	Paciente 577	2019-01-09 00:00:00.000	EN065	22+00 a la ronda de seguridad paciente duerme tranquilo con buen patrón de sue...
Notas_Evolucion	604	2022-05-08 13:44:00	Paciente 577	2019-01-10 00:00:00.000	EN065	00+00 se realiza ronda de enfermería encontrando paciente en cama con buen p...
Notas_Evolucion	605	2022-05-08 13:44:00	Paciente 577	2019-01-10 00:00:00.000	EN065	03+00 se realiza ronda de enfermería paciente duerme tranquilo con buen patrón ...
Notas_Evolucion	606	2022-05-08 13:44:00	Paciente 577	2019-01-10 00:00:00.000	AU555	VALORACIÓN DEL PACIENTE /SEGUIMIENTO DE ENFERMERÍA PACIENTE ...
Notas_Evolucion	607	2022-05-08 13:44:00	Paciente 577	2019-01-10 00:00:00.000	AU555	ADMINISTRACIÓN Y EDUCACIÓN DE MEDICAMENTOS Previa valoración i...
Notas_Evolucion	608	2022-05-08 13:44:00	Paciente 577	2019-01-10 00:00:00.000	EN065	05+30 se ofrece baño en ducha, paciente quien acepta pasa sin novedades, se r...
Notas_Evolucion	609	2022-05-08 13:44:00	Paciente 577	2019-01-10 00:00:00.000	AU440	07+00 recibo paciente de 85 años de edad en habitacion 202A con los sigui...

Recursos Literarios

Las notas de enfermería están compuestas por palabras y frases de alcance médico y para poder revisar si el vocabulario usado es correcto no, es necesario usar un recurso literario como base de conocimiento. Para este proyecto la principal fuente proviene es UMLS y SNOMEDCT-SP explicadas en la sección de Marco Conceptual. Adicionalmente se creó un diccionario [Ver tabla 5] de palabras y frases para normalizar abreviaciones y acrónimos (Ej. Dr.=Doctor, Efro=Enfermero, mg, mlg, mlgr=miligramos). Es importante señalar que entre más grande y de calidad sea el recurso lingüístico, mejor serán los resultados, por último, estos recursos son limitados en idioma español.

Tabla 5 - Recurso lingüístico con terminología médica en español

Origen	Conceptos	Contador
COMMON	440367	1204783
LOCAL CORPORA	13	1500
SNOMEDCTSP	6617	15748
UMLS	243790	379763

Limpieza, mapeo y enriquecimiento de datos

Siguiendo metodologías de ETL, los archivos originales se cargan tal como fueron recibidos a una zona denominada de staging área en el motor de base de dato;

Luego, estos empiezan un proceso de limpieza, dividido en dos partes y cada una con tres etapas, Pre-limpieza de datos con sus tres etapas (Raw, Pre-Clean y Clean) y Limpieza de datos con sus tres etapas (Raw, Pre-Clean y Clean).

Limpieza de datos

La tabla 6 describe los pasos requeridos para la limpieza de datos, básicamente se compone de dos rondas de limpieza. La primera ronda, denominada como fase de pre-limpieza parte de los datos crudos, cargados como llegaron de la fuente de datos, inicia con una revisión a nivel de caracteres (ASCII) donde se remueven caracteres innecesarios para el estudio, luego, mediante expresiones regulares, se arreglan palabras que pueden llegar pegadas con otras o con números, así mismo, se estandarizan siglas , horas y fechas, para terminar con un proceso de corrector ortográfico y re-ensamblado de los datos. El resultado de esta ronda son NN corregidas a nivel léxico para el idioma español y con unas etiquetas que permiten separar notas que contienen metainformación para la próxima ronda.

La segunda ronda de limpieza, parte de la salida de la ronda anterior, en esta, pueden llegar nuevos registros por nota, ya que la metainformación identificó más de una interacción entre enfermero-pacientes (Ver tabla 32 en Anexos), se remueven frases innecesarias (stop phrases) y palabras innecesarias (stopwords), se realiza nuevamente el proceso de limpieza como en el paso anterior, terminando con lematización de las palabras y re-ensamblado de datos.

Tabla 6 - Pre-Limpieza y limpieza de datos

PRE-Limpieza (Primera ronda de limpieza)	Limpieza (Segunda ronda de limpieza)
1A - Pre-Limpieza de datos (Pre-Clean)	**2A Limpieza de datos (Pre-Clean)**
Minusculas y Acentos	*Minusculas y Acentos*
ASCII - Caracteres de control	*ASCII - Caracteres de control*
ASCII - Caracteres Especiales	*ASCII - Caracteres Especiales*
ASCII - Caracteres para remover	*ASCII - Caracteres para remover*
ASCII - Caracteres tipo símbolo usados en las NN	*ASCII - Caracteres tipo símbolo usados en las NN*
ASCII - Caracteres Imprimibles	*ASCII - Caracteres Imprimibles*
Remover caracteres especiales duplicados consecutivamente	*Remover caracteres especiales duplicados consecutivamente*
Remover espacios duplicados consecutivamente	*Remover espacios duplicados consecutivamente*
	Remover Frases innecesarias
	Remover Stopwords
	Tokenización
1B - Pre-Limpieza de datos (Clean)	**2B - Limpieza de datos (Clean)**
UMLS	*UMLS*
Abreviaciones/Siglas/Acrónimos/Unidades de Medida	*Abreviaciones/Siglas/Acrónimos/Unidades de Medida*
Pre-Corrección ortográfica	*Pre-Corrección ortográfica*
Palabras con números y letras pegadas	*Palabras con números y letras pegadas*
Palabras con letras y números pegados	*Palabras con letras y números pegados*
Agregar espacios a caracteres especiales mediante patrones	*Agregar espacios a caracteres especiales mediante patrones*
Fechas/Horas dentro de las NN	*Fechas/Horas dentro de las NN*
Corrección ortográfica	*Corrección ortográfica*
	Lematizacion
1C - Pre-Limpieza de datos (Reassembled)	**2C - Limpieza de datos (Reassembled)**
Re-ensamblar los datos	*Re-ensamblar los datos*

Las NN cargadas en la herramienta se mantienen en original y se crean dos nuevas columnas a partir de la original, la primera columna llamada Pre-Clean recibe los primeros cambios de la pre-limpieza de datos. La segunda columna llamada Clean es la etapa final de la pre-limpieza de datos, los datos han recibido corrección ortográfica más otras técnicas de limpieza, mapeo y enriquecimiento de datos.

1A - Pre-Limpieza de datos (Pre-Clean)

Minúsculas y Acentos

Los primeros cambios realizados sobre las NN son el de remover vocales acentuadas por vocales sin acento, y todo el texto es convertido a minúscula. Se mantiene la letra ñ.

ASCII - Caracteres de control

Los caracteres de control según el código ASCII son removidos de las NN y se reemplazan por el carácter pipe (|), este será usado para saltos de línea en pasos posteriores. Un ejemplo son las separaciones mediante tabuladores, viñetas, saltos de Línea y otros caracteres. (Ver tabla 7)

Tabla 7 - Caracteres de Control según ASCII / ISO 8859-1

Dec	Value	Description	TREATMENT
0	^@	Null (NUL)	CONTROL CHARACTERES
1	^A	Start of heading (SOH)	CONTROL CHARACTERES
2	^B	Start of text (STX)	CONTROL CHARACTERES
3	^C	End of text (ETX)	CONTROL CHARACTERES
4	^D	End of transmission (EOT)	CONTROL CHARACTERES
5	^E	Enquiry (ENQ)	CONTROL CHARACTERES
6	^F	Acknowledge (ACK)	CONTROL CHARACTERES
7	^G	Bell (BEL)	CONTROL CHARACTERES
8	^H	Backspace (BS)	CONTROL CHARACTERES
9	^I	Horizontal tab (HT)	CONTROL CHARACTERES
10	^J	Line feed (LF)	CONTROL CHARACTERES
11	^K	Vertical tab (VT)	CONTROL CHARACTERES
12	^L	New page/form feed (FF)	CONTROL CHARACTERES
13	^M	Carriage return (CR)	CONTROL CHARACTERES
14	^N	Shift out (SO)	CONTROL CHARACTERES
15	^O	Shift in (SI)	CONTROL CHARACTERES
16	^P	Data link escape (DLE)	CONTROL CHARACTERES
17	^Q	Device control 1 (DC1)	CONTROL CHARACTERES
18	^R	Device control 2 (DC2)	CONTROL CHARACTERES
19	^S	Device control 3 (DC3)	CONTROL CHARACTERES
20	^T	Device control 4 (DC4)	CONTROL CHARACTERES
21	^U	Negative acknowledge (NAK)	CONTROL CHARACTERES
22	^V	Synchronous idle (SYN)	CONTROL CHARACTERES
23	^W	End of transmission block (ETB)	CONTROL CHARACTERES
24	^X	Cancel (CAN)	CONTROL CHARACTERES
25	^Y	End of medium (EM)	CONTROL CHARACTERES
26	^Z	Substitute (SUB)	CONTROL CHARACTERES
27	^[	Escape (ESC)	CONTROL CHARACTERES
28	^\	File separator (FS)	CONTROL CHARACTERES
29	^]	Group separator (GS)	CONTROL CHARACTERES
30	^^	Record separator (RS)	CONTROL CHARACTERES
31	^_	Unit separator (US)	CONTROL CHARACTERES
127	DEL	Delete (DEL)	CONTROL CHARACTERES

ASCII - Caracteres Especiales

Son caracteres que tienen un uso especial sobre las NN, no se ejecuta ninguna acción sobre ellos. (Ver tabla 8)

Tabla 8 - Caracteres Especiales según ASCII / ISO 8859-1

Dec	Hex	Value	Description	TREATMENT
32	20		Space	SPECIAL CHARACTERS
43	2B	+	Plus sign	SPECIAL CHARACTERS
45	2D	-	Hyphen/Minus	SPECIAL CHARACTERS
46	2E	.	Full stop/Period	SPECIAL CHARACTERS
47	2F	/	Solidus/Slash	SPECIAL CHARACTERS
58	3A	:	Colon	SPECIAL CHARACTERS

ASCII - Caracteres para remover

En su mayoría son caracteres extendidos del estándar ASCII, son removidos de las NN, ya que no aportan utilidad en el estudio y si interfieren en algunos procesos.

ASCII - Caracteres tipo símbolo usados en las NN

En su mayoría son caracteres tipo símbolos usados en las NN, se agrega un espacio entre ellos para facilitar su manipulación. (Ver tabla 9)

Tabla 9 - Caracteres tipo símbolo según ASCII / ISO 8859-1

Dec	Hex	Value	Description	TREATMENT
35	23	#	Number sign	SPACE ON BOTH SIDES
36	24	$	Dollar sign	SPACE ON BOTH SIDES
37	25	%	Percent sign	SPACE ON BOTH SIDES
40	28	(	Left parenthesis	SPACE ON BOTH SIDES
41	29	)	Right parenthesis	SPACE ON BOTH SIDES
44	2C	,	Comma	SPACE ON BOTH SIDES
59	3B	;	Semicolon	SPACE ON BOTH SIDES
60	3C	<	Less-than sign	SPACE ON BOTH SIDES
61	3D	=	Equal/Equality sign	SPACE ON BOTH SIDES
62	3E	>	Greater-than sign	SPACE ON BOTH SIDES
91	5B	[	Left square bracket	SPACE ON BOTH SIDES
93	5D	]	Right square bracket	SPACE ON BOTH SIDES
123	7B	{	Left curly bracket	SPACE ON BOTH SIDES
125	7D	}	Right curly bracket	SPACE ON BOTH SIDES

ASCII - Caracteres Imprimibles

Son los caracteres usados en el alfabeto español y sin requeridos en las notas de enfermería, no hay ninguna acción sobre este tipo de caracteres. (Ver tabla 10).

Tabla 10 - Caracteres imprimibles según ASCII / ISO 8859-1

Dec	Value	Description	TREATMENT
38	&	Ampersand	PRINTABLE CHARACTERS
42	*	Asterisk	PRINTABLE CHARACTERS
48	0	Digit zero	PRINTABLE CHARACTERS
49	1	Digit one	PRINTABLE CHARACTERS
50	2	Digit two	PRINTABLE CHARACTERS
51	3	Digit three	PRINTABLE CHARACTERS
52	4	Digit four	PRINTABLE CHARACTERS
53	5	Digit five	PRINTABLE CHARACTERS
54	6	Digit six	PRINTABLE CHARACTERS
55	7	Digit seven	PRINTABLE CHARACTERS
56	8	Digit eight	PRINTABLE CHARACTERS
57	9	Digit nine	PRINTABLE CHARACTERS
64	@	Commercial at/At sign	PRINTABLE CHARACTERS
65	A	Latin capital letter A	PRINTABLE CHARACTERS
66	B	Latin capital letter B	PRINTABLE CHARACTERS
67	C	Latin capital letter C	PRINTABLE CHARACTERS
68	D	Latin capital letter D	PRINTABLE CHARACTERS

Dec	Value	Description	TREATMENT
69	E	Latin capital letter E	PRINTABLE CHARACTERS
70	F	Latin capital letter F	PRINTABLE CHARACTERS
71	G	Latin capital letter G	PRINTABLE CHARACTERS
72	H	Latin capital letter H	PRINTABLE CHARACTERS
73	I	Latin capital letter I	PRINTABLE CHARACTERS
74	J	Latin capital letter J	PRINTABLE CHARACTERS
75	K	Latin capital letter K	PRINTABLE CHARACTERS
76	L	Latin capital letter L	PRINTABLE CHARACTERS
77	M	Latin capital letter M	PRINTABLE CHARACTERS
78	N	Latin capital letter N	PRINTABLE CHARACTERS
79	O	Latin capital letter O	PRINTABLE CHARACTERS
80	P	Latin capital letter P	PRINTABLE CHARACTERS
81	Q	Latin capital letter Q	PRINTABLE CHARACTERS
82	R	Latin capital letter R	PRINTABLE CHARACTERS
83	S	Latin capital letter S	PRINTABLE CHARACTERS
84	T	Latin capital letter T	PRINTABLE CHARACTERS
85	U	Latin capital letter U	PRINTABLE CHARACTERS
86	V	Latin capital letter V	PRINTABLE CHARACTERS

Dec	Value	Description	TREATMENT
87	W	Latin capital letter W	PRINTABLE CHARACTERS
88	X	Latin capital letter X	PRINTABLE CHARACTERS
89	Y	Latin capital letter Y	PRINTABLE CHARACTERS
90	Z	Latin capital letter Z	PRINTABLE CHARACTERS
95	_	Underscore/Low line	PRINTABLE CHARACTERS
97	a	Latin small letter a	PRINTABLE CHARACTERS
98	b	Latin small letter b	PRINTABLE CHARACTERS
99	c	Latin small letter c	PRINTABLE CHARACTERS
100	d	Latin small letter d	PRINTABLE CHARACTERS
101	e	Latin small letter e	PRINTABLE CHARACTERS
102	f	Latin small letter f	PRINTABLE CHARACTERS
103	g	Latin small letter g	PRINTABLE CHARACTERS
104	h	Latin small letter h	PRINTABLE CHARACTERS
105	i	Latin small letter i	PRINTABLE CHARACTERS
106	j	Latin small letter j	PRINTABLE CHARACTERS
107	k	Latin small letter k	PRINTABLE CHARACTERS
108	l	Latin small letter l	PRINTABLE CHARACTERS
109	m	Latin small letter m	PRINTABLE CHARACTERS

Dec	Value	Description	TREATMENT
110	n	Latin small letter n	PRINTABLE CHARACTERS
111	o	Latin small letter o	PRINTABLE CHARACTERS
112	p	Latin small letter p	PRINTABLE CHARACTERS
113	q	Latin small letter q	PRINTABLE CHARACTERS
114	r	Latin small letter r	PRINTABLE CHARACTERS
115	s	Latin small letter s	PRINTABLE CHARACTERS
116	t	Latin small letter t	PRINTABLE CHARACTERS
117	u	Latin small letter u	PRINTABLE CHARACTERS
118	v	Latin small letter v	PRINTABLE CHARACTERS
119	w	Latin small letter w	PRINTABLE CHARACTERS
120	x	Latin small letter x	PRINTABLE CHARACTERS
121	y	Latin small letter y	PRINTABLE CHARACTERS
122	z	Latin small letter z	PRINTABLE CHARACTERS
124	|	Vertical line/Vertical bar	PRINTABLE CHARACTERS
176	°	Degree sign	PRINTABLE CHARACTERS
209	Ñ	Latin capital letter N with tilde	PRINTABLE CHARACTERS
241	ñ	Latin small letter n with tilde	PRINTABLE CHARACTERS

Remover caracteres especiales duplicados consecutivamente

Son caracteres especiales y símbolos que se suplican consecutivamente más de dos veces, tales como “++++++++”, “********”, “--------”,”*_*_*_*_”, entre otros más. Son reemplazados por una sola repetición.

Remover espacios duplicados consecutivamente

Es una limpieza sobre el carácter de espacio, si se detecta este carácter repetido consecutivamente muchas veces, se reemplaza por tan solo una repetición.

Tokenización

Se realiza Tokenización básica de los datos usando como carácter de espacio como separador, los datos divididos se consolidan en una sola tabla, la cual mantiene el linaje de los datos. Este proceso no remueve stopwords.

Bolsa de Palabras

Se crea una bolsa de palabras, mediante la generación de una tabla de valores agregados, sobre esta tabla se calcula la frecuencia y participación de las palabras sobre el total de las palabras.

1B - Pre-Limpieza de datos (Clean)

UMLS

Cada palabra de la bolsa de palabras es comparada contra la base de datos de UMLS con el fin de validar si es pertenece a alguna jerga médica. Si aparece en este diccionario, la palabra es marcada como tal y se excluye de las revisiones posteriores.

Abreviaciones/Siglas/Acrónimos/Unidades de Medida

En este paso se usa el diccionario de palabras creado en la aplicación y se normaliza aquellas palabras que son abreviaciones, siglas, acrónimos o unidades de medida. Con esto, se estandariza el lenguaje, permitiendo búsquedas más precisas en la siguiente etapa. (Ver tabla 11).

Tabla 11 - Muestra de abreviaciones en las NN

Mapping_Key	Mapping_Language_Key	Mapping_Language_Value	Mapping_Type_Key	Mapping_Type_Value	Mapping_Action_Key	Mapping_Action_Value	Mapping_Value	Mapping_Value_Description	Mapping_Len_Value	Mapping_Number_Tokens	Mapping_Fixed_Value
905	1	español	1	MAESTRO	3	Cambiar	cc/cd	cc/cd	5	1	cc/cd
10	1	español	1	MAESTRO	3	Cambiar	cc/kg	centímetro cubico por kilogramo	5	1	CC/KG
902	1	español	1	MAESTRO	3	Cambiar	cc/gr	cc/gr	5	1	cc/gr
22	1	español	1	MAESTRO	3	Cambiar	gr/dl	gramos por decilitro	5	1	GR/DL
21	1	español	1	MAESTRO	3	Cambiar	gr/ml	gramo por mililitro	5	1	GR/ML
31	1	español	1	MAESTRO	3	Cambiar	m/seg	metros por segundo	5	1	M/SEG
34	1	español	1	MAESTRO	3	Cambiar	mcg/h	microgramos por hora	5	1	MCG/H
41	1	español	1	MAESTRO	3	Cambiar	meq/h	miliequivalentes por hora	5	1	MEQ/H
24	1	español	1	MAESTRO	3	Cambiar	kg/dl	kilogramo por decilitro	5	1	KG/DL
47	1	español	1	MAESTRO	3	Cambiar	mg/gr	miligramos por gramo	5	1	MG/GR
49	1	español	1	MAESTRO	3	Cambiar	mg/kg	miligramos por kilogramo	5	1	MG/KG
42	1	español	1	MAESTRO	3	Cambiar	mg/ml	miligramo por mililitro	5	1	MG/ML
52	1	español	1	MAESTRO	3	Cambiar	ml/dl	mililitro por decilitro	5	1	ML/DL
900	1	español	1	MAESTRO	3	Cambiar	mm/hg	mm/hg	5	1	mm/hg
45	1	español	1	MAESTRO	3	Cambiar	mg/dl	miligramos por decilitro	5	1	MG/DL
61	1	español	1	MAESTRO	3	Cambiar	ui/kg	unidades por kilogramo	5	1	UI/KG
58	1	español	1	MAESTRO	3	Cambiar	ui/cc	unidades por centímetro cubico	5	1	UI/CC

Pre-Corrección ortográfica

Cada palabra de la bolsa de palabras, excluyendo las encontradas en UMLS, son revisadas mediante un diccionario creado manualmente en el sistema, esto permite hacer exclusiones de palabras regionales o modismos que no se encuentran en otra parte, o también forzar las correcciones de palabras que no son bien arregladas por el corrector ortográfico automático. (Ver tabla 12).

Tabla 12 - Muestra de palabras para corregir ortografía manualmente

Mapping_Key	Mapping_Language_Key	Mapping_Language_Value	Mapping_Type_Key	Mapping_Type_Value	Mapping_Action_Key	Mapping_Action_Value	Mapping_Value	Mapping_Value_Description	Mapping_Len_Value	Mapping_Number_Tokens	Mapping_Fixed_Value
3626	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemiodinamicamente		18	1	hemodinamicamente
3627	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemmodinamicamente		18	1	hemodinamicamente
3630	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodianamicamente		18	1	hemodinamicamente
3631	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodibnamicamente		18	1	hemodinamicamente
3633	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodicnamicamente		18	1	hemodinamicamente
3635	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodimnamicamente		18	1	hemodinamicamente
3638	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodinaimicamente		18	1	hemodinamicamente
3641	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodinameicamente		18	1	hemodinamicamente
3642	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodinamiacamente		18	1	hemodinamicamente
3645	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodinamicaamente		18	1	hemodinamicamente
3651	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodinamicamenete		18	1	hemodinamicamente
3652	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodinamicamenmte		18	1	hemodinamicamente
3656	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodinamicamenter		18	1	hemodinamicamente
3657	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodinamicamentes		18	1	hemodinamicamente
3659	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodinamicamentre		18	1	hemodinamicamente
3660	1	español	78	PRE-SPELLCHECK	3	Cambiar	hemodinamicamentte		18	1	hemodinamicamente

Palabras con números y letras pegadas

En el análisis inicial de los datos, se encontró palabras como 100cc, 100litros, 100gramos, cuyo problema es el no tener un separador como el carácter de espacio y que no permite que la palabra esté bien formada. Mediante expresiones regulares, se encuentra este patrón y se agrega un espacio entre la palabra.

Palabras con letras y números pegados

En el análisis inicial de los datos, se encontró palabras como mq024501, cut33000566, lacosamida200mg, cuyo problema es el no tener un separador como el carácter de espacio y que no permite que la palabra esté bien formada. Mediante expresiones regulares, se encuentra este patrón y se agrega un espacio entre la palabra.

Agregar espacios a caracteres especiales mediante patrones

En el Paso 1 de la Pre-limpieza de datos se excluyeron caracteres especiales como (+,-,/,:,.), en este segundo paso, hay palabras usando estos caracteres, pero que están mal formadas gramaticalmente. Mediante expresiones regulares se encuentran patrones en prefijos, infijos y sufijos en palabras y se arreglan.

Fechas/Horas dentro de las NN

En el análisis de datos se detectó que un solo registro de las NN podría tener embebida varias interacciones de los profesionales de enfermería con el paciente en diferentes horas. Mediante expresiones regulares, se encuentran patrones con formato de fecha/hora en prefijos, infijos y sufijos de las palabras y se resaltan con el formato {hh:mm} o {aaaa/mm/dd}. En procesos posteriores, está hora será usada para dividir la frase.

Corrección ortográfica

Cada palabra de la bolsa de palabras, excluyendo las encontradas las revisiones anteriores, son revisadas mediante un corrector automático de ortografía en Python para el idioma español, esto permite arreglar palabras mal escritas, un ejemplo de la potencia que tiene este corrector es la de poder arreglar la palabra paciente, la cual se encontró mal escrita hasta 177 veces. La ilustración 6 ilustra el código Python usado pata la corrección ortográfica. Esta revisión no excluye stopwords.

```python
import pyodbc
import language_tool_python
from datetime import datetime

server = '[illegible]'
database = 'NursingNotes'
username = '[illegible]'
password = '[illegible]'
conn = pyodbc.connect('DRIVER={ODBC Driver 17 for SQL Server};SERVER='+server+';DATABASE='+database+';UID='+username+';PWD='+ password)
cursor = conn.cursor()

tool = language_tool_python.LanguageTool('es-ES')

#Select Query
print ('Reading data from table')
tsql = "SELECT GroupByToken_Key,Token_Value_Pre_Fixed FROM STG.GroupByToken WHERE Token_Value_Fixed IS NULL"
tsqlu = "UPDATE STG.GroupByToken SET Token_Value_Fixed_Type=4, Token_Value_Fixed = ? WHERE GroupByToken_Key= ? "

cursor.execute(tsql)
rows = cursor.fetchall()
start_date = datetime.now()
print('Starting Update at ',start_date)
for fila in rows:
    spellcheck=language_tool_python.utils.correct(fila.Token_Value_Pre_Fixed,tool.check(fila.Token_Value_Pre_Fixed))
    key=fila.GroupByToken_Key
    with cursor.execute(tsqlu,spellcheck,key):
        pass
end_date = datetime.now()
print('Finishing Update at ',end_date)
duracion = (end_date - start_date).total_seconds() / 60.0
print('language_tool tardo {} minutos'.format(duracion))
#language_tool tardo 60.372363566666664 minutos
```

Ilustración 6 - Código Python para corrección ortográfica en español

1C - Pre-Limpieza de datos (Reassembled)

Este proceso reconstruye la frase original, pero con las palabras pre-limpias, y es la etapa final de la Pre-limpieza de datos.

Data Split

Con los datos limpios del proceso anterior, se procede a crear una nueva tabla consolidada con las notas, usando como fuente de información las Notas_Clean, adicionalmente, este proceso genera nueva metada con las interacciones ocultas de los profesionales de enfermería y que estaban inmersas dentro de las notas originales. Adicionalmente, se generan nuevas frases a partir de los caracteres de control que estaban identificados bajo el carácter pipe (|).

Limpieza de Datos

Después de crear una tabla con nuevas frases según el paso anterior, se aplican pasos similares a los del proceso de Pre-limpieza, con algunas variaciones, así mismo, se ejecutan nuevas tareas de limpieza con el objetivo de preparar los datos para ser usados con los algoritmos de NLP, detallados en los siguientes numerales.

2A - Limpieza de datos (*Primera ronda de limpieza*)

Este segundo proceso de limpieza se ejecuta sobre las Notas_Pre-Clean, en esta etapa los datos ya tienen cierto nivel de limpieza, sin embargo, requieren repetir algunos pasos de limpieza, similares a los realizados a los de Pre-Limpieza, también es posible incluir nuevas limpiezas, que antes no eran posibles.

- Minúsculas y tildes (Clean)
- ASCII - Caracteres de control, especiales, para remover, símbolos e imprimibles (Clean)
- Remover caracteres especiales duplicados consecutivamente (Clean)
- Remover espacios duplicados consecutivamente (Clean)

Los siguientes pasos cambian un poco o son nuevos, con respecto a la Pre-Limpieza de datos

Remover Frases innecesarias

En esta parte del proceso, se remueven frases que previamente se han identificado como comodines y que no aportan contexto en el análisis de las notas de enfermería. Se diseñó una tabla para ser usada en la parametrización de frases que no se requieren analizar, así mismo, se pueden activar o desactivar frases (Ver tabla 13). La aplicación soporta frases en otros lenguajes. Las frases innecesarias fueron revisadas en trabajo conjunto con los PDE. (Ver ejemplos en marco conceptual)

Tabla 13 - Frases a excluir

StopWords_key	StopWords_Value	StopWords_Language	StopWords_Type	StopWords_Active	StopWords_DataLength	StopWords_Quantity	StopWords_Notes	StopWords_Load_DT
277	actividades de enfermeria	ESP	Phrase	1	25	3	NULL	2021-10-13 10:49:00.283
278	administracion de medicamentos	ESP	Phrase	1	30	3	NULL	2021-10-13 10:49:00.283
279	asistencia en necesidades basicas	ESP	Phrase	1	33	4	NULL	2021-10-13 10:49:00.283
280	asistir necesidades basicas	ESP	Phrase	1	27	3	NULL	2021-10-13 10:49:00.283
281	control de liquidos administrados y eliminados	ESP	Phrase	1	46	6	NULL	2021-10-13 10:49:00.283
282	control de signos vitales	ESP	Phrase	1	25	4	NULL	2021-10-13 10:49:00.283
283	control y registro de signos vitales	ESP	Phrase	1	36	6	NULL	2021-10-13 10:49:00.283
284	diagnostico de enfermeria	ESP	Phrase	1	25	3	NULL	2021-10-13 10:49:00.283
285	finalmente se da espacio para hacer pregunta y resolver dudas	ESP	Phrase	1	61	10	NULL	2021-10-13 10:49:00.283
286	informacion de ingreso	ESP	Phrase	1	22	3	NULL	2021-10-13 10:49:00.283
287	informar cambios	ESP	Phrase	1	16	2	NULL	2021-10-13 10:49:00.283
288	ingreso hospitalizacion por enfermeria	ESP	Phrase	1	38	4	NULL	2021-10-13 10:49:00.283
289	intervencion inmediata	ESP	Phrase	1	22	2	NULL	2021-10-13 10:49:00.283
290	momento de cuidado	ESP	Phrase	1	18	3	NULL	2021-10-13 10:49:00.283
291	paciente despierto alerta y tranquilo	ESP	Phrase	1	37	5	NULL	2021-10-13 10:49:00.283
292	plan de enfermeria	ESP	Phrase	1	18	3	NULL	2021-10-13 10:49:00.283
293	politraumatismo	ESP	Phrase	1	15	1	NULL	2021-10-13 10:49:00.283
294	se abre folio para medio de contraste	ESP	Phrase	1	37	7	NULL	2021-10-13 10:49:00.283
295	se abre folio para registro de contraste para tac	ESP	Phrase	1	49	9	NULL	2021-10-13 10:49:00.283
296	se brinda espacio para hacer preguntas y aclarar dudas	ESP	Phrase	1	54	9	NULL	2021-10-13 10:49:00.283
297	se educa en cuidado y autocuidado	ESP	Phrase	1	33	6	NULL	2021-10-13 10:49:00.283
298	se evalua efecto terapeutico	ESP	Phrase	1	28	4	NULL	2021-10-13 10:49:00.283
299	se realiza protocolo de bienvenida	ESP	Phrase	1	34	5	NULL	2021-10-13 10:49:00.283
300	sin aparente deficit neurologico afebril	ESP	Phrase	1	40	5	NULL	2021-10-13 10:49:00.283
301	sin complicaciones durante la ejecucion	ESP	Phrase	1	39	5	NULL	2021-10-13 10:49:00.283
302	valoracion e identificacion de riesgos	ESP	Phrase	1	38	5	NULL	2021-10-13 10:49:00.283

Remover Stopwords

En este proceso se remueven pronombres, artículos y palabras que no aportan significado lingüístico a la oración, para ello, se construyó un catálogo donde se pueden agregar nuevas palabras (Ver tabla 14), así mismo, se pueden activar o desactivar palabras. La aplicación soporta palabras en otros lenguajes. (Ver ejemplos en marco conceptual). Para la construcción de los stopwords se usó diccionarios y librerías de Python en español para construir la base de conocimiento (KB).

Tabla 14 - Stopwords

StopWords_key	StopWords_Value	StopWords_Language	StopWords_Type	StopWords_Active	StopWords_DataLength	StopWords_Quantity	StopWords_Notes	StopWords_Load_DT
1	a causa de	ESP	connectors	1	10	3	NULL	2021-10-13 11:25:14.110
2	a consecuencia de	ESP	connectors	1	17	3	NULL	2021-10-13 11:25:14.110
3	a fin de	ESP	connectors	1	8	3	NULL	2021-10-13 11:25:14.110
4	a fin de que	ESP	connectors	1	12	4	NULL	2021-10-13 11:25:14.110
5	a menos que	ESP	connectors	1	11	3	NULL	2021-10-13 11:25:14.110
6	a partir de ahi	ESP	connectors	1	15	4	NULL	2021-10-13 11:25:14.110
7	a pesar de	ESP	connectors	1	10	3	NULL	2021-10-13 11:25:14.110
8	a proposito de	ESP	connectors	1	14	3	NULL	2021-10-13 11:25:14.110
9	a saber	ESP	connectors	1	7	2	NULL	2021-10-13 11:25:14.110
10	abril	ESP	months	1	5	1	NULL	2021-10-13 11:25:14.110
11	aca	ESP	pronouns	1	3	1	NULL	2021-10-13 11:25:14.110
12	actualmente	ESP	connectors	1	11	1	NULL	2021-10-13 11:25:14.110
13	ademas	ESP	connectors	1	6	1	NULL	2021-10-13 11:25:14.110
14	agosto	ESP	months	1	6	1	NULL	2021-10-13 11:25:14.110
15	ahi	ESP	pronouns	1	3	1	NULL	2021-10-13 11:25:14.110
16	ahora	ESP	connectors	1	5	1	NULL	2021-10-13 11:25:14.110
17	ahora que	ESP	connectors	1	9	2	NULL	2021-10-13 11:25:14.110
18	al contrario	ESP	connectors	1	12	2	NULL	2021-10-13 11:25:14.110
19	al final	ESP	connectors	1	8	2	NULL	2021-10-13 11:25:14.110
20	al principio	ESP	connectors	1	12	2	NULL	2021-10-13 11:25:14.110

Tokenización

Se realiza Tokenización básica de los datos usando el carácter de espacio como separador, los datos divididos se consolidan en una sola tabla, la cual mantiene el linaje de los datos. Este proceso remueve stopwords.

Bolsa de Palabras

Se crea una bolsa de palabras (BoW), mediante la generación de una tabla de valores agregados, sobre esta tabla se calcula la frecuencia y participación de las palabras sobre el total de las palabras. No hay stop words en esta etapa.

Lematización

Este paso lematiza cada palabra a su raíz, simplificando el lenguaje para ser analizado por algoritmos de NLP y para facilitar consultas sobre los datos. La ilustración 7 muestra una rutina en Python que se conecta a la base de datos, tomas los tokens de la bolsa de palabras (BoW), las lematiza y actualiza en una columna nueva en la BoW, con ello se mantiene la palabra original y la lematizada para futuras comparaciones. La lematización es un paso importante, ya que reduce la bolsa de palabras y mejora la predictibilidad de los algoritmos aplicados en las etapas siguientes, porque se reduce el número de palabras a procesar. Comparando la BoW original vs la lematizada, se redujo en un 46% el número de tokens únicos a procesar. Ver código implementado en ilustración 11.

```
#pip install pyodbc
import pyodbc
import spacy
from datetime import datetime
nlp = spacy.load('es_core_news_lg')
server = '[illegible]'
database = 'NursingNotes'
username = '[illegible]'
password = '[illegible]'
conn = pyodbc.connect('DRIVER={ODBC Driver 17 for SQL Server};SERVER='+server+';DATABASE='+database+';UID='+username+';PWD='+ password)
cursor = conn.cursor()

#Select Query
print ('Reading data from table')
tsql = "SELECT GroupByToken_Key,Token_Value_Pre_Fixed FROM STG.GroupByToken WHERE Token_Value_Fixed IS NULL"
tsqlu = "UPDATE STG.GroupByToken SET Token_Value_Fixed_Type=4, Token_Value_Fixed = ? WHERE GroupByToken_Key= ? "
cursor.execute(tsql)
rows = cursor.fetchall()
start_date = datetime.now()
print('Starting Update at ',start_date)

for fila in rows:
    lemma=(" ".join([token.lemma_ for token in nlp(fila.Token_Value_Pre_Fixed)]))
    key=fila.GroupByToken_Key
    with cursor.execute(tsqlu,lemma,key):
        pass
end_date = datetime.now()
print('Finishing Update at ',end_date)
duracion = (end_date - start_date).total_seconds() / 60.0
print('Lemmatization tardo {} minutos'.format(duracion))
#Lemmatization tardo 27.7575481 minutos
```

Ilustración 7 - Código Python para Lematización en español

2B - Limpieza de datos (Segunda ronda de limpieza)

- UMLS (Clean)
- Abreviaciones/Siglas/Acrónimos/Unidades de Medida (Clean)
- Pre-Corrección ortográfica (Clean)
- Palabras con números/letras y letras/números pegados (Clean)
- Agregar espacios a caracteres especiales mediante patrones (Clean)
- Fechas/Horas dentro de las NN (Clean)
- Corrección ortográfica (Clean)

2C - Limpieza de datos (Reassembled)

Este proceso reconstruye la frase original, pero con las palabras limpias, sin frases innecesarias, sin stopwords y con el lema de cada palabra. Es la etapa final de la Limpieza de datos. (Ver tabla 15)

Tabla 15 - Muestra de Limpieza de datos (Raw, Pre-Clean y Clean) en las NN

Clean_Notes_Key	Clean_Notes_SRC_Table_Key	Clean_Notes_SRC_Key	Clean_Notes_Date	Clean_Notes_Time	Note_ORI	Note_PRE-CLEAN	Note_CLEAN	Upload_DT
1	1	2	2019-01-20	07:34:23.0000000	cambios de posición	cambios de posición	cambio posición	2021-12-28 21:42:00
2	1	3	2019-01-20	07:39:57.0000000	1 buretrol	1 buretrol	1 buretrol	2021-12-28 21:42:00
3	1	2	2019-01-20	07:34:23.0000000	lubricación de piel	lubricación de piel	lubricación piel	2021-12-28 21:42:00
4	1	514	2019-01-29	22:00:00.0000000	{22:00} alprazolam 0.5 mg por sonda orogástrico	alprazolam 0.5 mg por sonda orogástrico	alprazolam 0.5 mg por sonda orogástrico	2021-12-28 21:42:00
5	1	436	2019-01-10	01:00:00.0000000	{01:00} Se realiza ronda de enfermería encontrando ...	Se realiza ronda de enfermería encontrando Pacient...	realizar ronda enfermería encontrar Paciente cama co...	2021-12-28 21:42:00
6	1	1	2019-01-20	05:00:00.0000000	{05:00} Paciente quien continúa en unidad de cuidad...	Paciente quien continúa en unidad de cuidados inten...	Paciente quien continúar unidad cuidado intensivo con...	2021-12-28 21:42:00
7	1	1	2019-01-20	07:00:12.0000000	queda Paciente en unidad de cuidados intensivos co...	queda Paciente en unidad de cuidados intensivos co...	quedar Paciente unidad cuidado intensivo con Medida...	2021-12-28 21:42:00
8	1	190	2019-01-03	08:17:25.0000000	Se solicita AL servicio de Farmacia 2 tiras y 2 lanceta...	Se solicita AL servicio de Farmacia 2 tiras y 2 lanceta...	solicitar al servicio Farmacia 2 tira 2 lanceta tomar gluc...	2021-12-28 21:42:00
9	1	4	2019-01-20	09:54:29.0000000	Se realiza cambio de posición más lubricación de piel ...	Se realiza cambio de posición más lubricación de piel ...	realizar cambio posición más lubricación piel asistir nec...	2021-12-28 21:42:00
10	1	3	2019-01-20	07:39:57.0000000	según protocolo institucional	según protocolo institucional	protocolo institucional	2021-12-28 21:42:00
11	1	439	2019-01-10	06:05:13.0000000	Administración y educación de Medicamentos	Administración y educación de Medicamentos	Administración educación Medicamentos	2021-12-28 21:42:00

HERRAMIENTAS USADAS

SQL Server

Para la persistencia de los datos, se eligió motor de base de datos relacional (RDBM) SQL server 2019 64 bits Developer Edition, con colación (collation) SQL_Latin1_General_CP1_CI_AI. Está elección permite persistir los datos cargados desde los archivos de Excel y a partir de ellos, crear un linaje de datos en las diferentes etapas de pre y procesamiento de los datos, así como almacenar las tablas paramétricas requeridas a lo largo de todo el proceso de transformación de datos. Se utilizó esta herramienta pensando en una fácil implementación en la infraestructura tecnológica de la CUS. Por último, el collate seleccionado facilita trabajar con mayúsculas y minúsculas, lo mismo con palabras con acentos y sin acentos (Ej. PALABRA=palabra, último=último). El licenciamiento de esta herramienta es paga para entornos productivos, la versión usada para este proyecto es para desarrolladores.

Python

Para las funciones avanzadas de analítica de texto, se convino el poder del motor de base de datos, con las características que ofrece Python. Python es un lenguaje interpretado de programación de alto nivel y código abierto. La versión usada es 3.10.4 y las librerías usadas son pyodbc (Para la conexión contra la base de datos), *spacy* (Para la lematización de los datos en español, *nltk* es otra opción, pero los algoritmos para el idioma español son limitados), SpellCheck (para la revisión ortográfica).

Power BI

Es una de las herramientas de inteligencia de negocios más usadas en el mercado para la visualización de datos a través de paneles, informes y reportes, la cual se puede conectar a múltiples fuentes de información, en especial con bases de datos tipo SQL Server y la última versión (2.105.664.0 64-bit - May 2022) ya incluye características especiales para análisis de texto, ligadas al tipo de licenciamiento contratado. Se eligió esta herramienta, ya que pertenece al ecosistema Microsoft usado por la CUS, el licenciamiento es híbrido, algunas funcionalidades se pueden usar sin ningún pago, algunas otras requieren una licencia especial.

MICMAC

Es una herramienta de análisis estructural y ofrece la posibilidad de describir un sistema con ayuda de una matriz que relaciona todos sus elementos constitutivos. Con la herramienta y siguiendo la metodología de la prospectiva, resaltar las variables influyentes y dependientes esenciales para la evolución del sistema. El análisis estructural se realiza por un grupo de trabajo compuesto por actores y expertos con experiencia demostrada, en este caso los profesionales de enfermería. Las diferentes fases del método son listar las variables clave, la revisión de relaciones entre variables y la revisión de variables clave. [31]

Erwin data modeler

Es un software para el modelado de datos, con ella, es posible generar los modelos lógicos y físicos en una arquitectura de datos, para luego ser llevados a una base de datos, también facilita en un alto nivel revisar las relaciones entre todas las entidades, estandarizar la mnemotecnia usada y evitar redundancia de datos. [32]

RESULTADOS

Análisis descriptivo

En la revisión léxica de los datos, se analizó cómo está construido el corpus de las notas, se detectó:

- Frases duplicadas
- No hay uniformidad en la redacción, puesto que son varios los enfermeros que ingresan registros.
- Errores de ortografía y digitación.
- Uso de plantillas para el diligenciamiento de las anotaciones.
- Un solo registro puede tener embebidas más interacciones enfermero-paciente.
- Hay varios registros con demasiado texto, con caracteres espaciales tales como tabuladores, viñetas, saltos de línea.
- Falta de estandarización de acrónimos, sinónimos y abreviaciones.

Análisis cualitativo

Las notas son texto libre no estructurado de tipo cualitativo elaborados por los PDE, la tabla 16 y 17 resume la cantidad de registros por fuente y estadísticas básicas del total de los datos.

Tabla 16 - Contador de frases totales, únicas y vacías en las NN

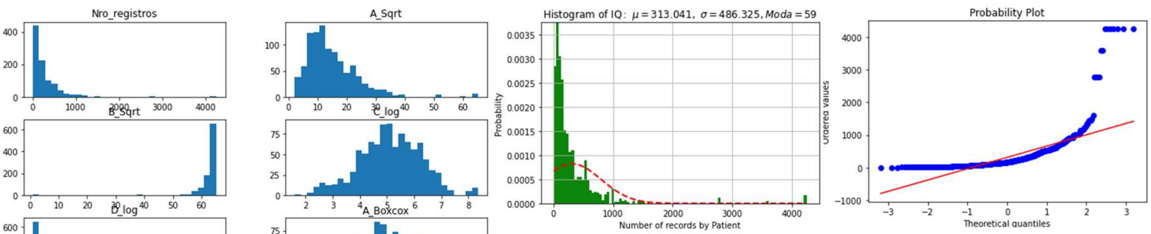
Origen	FECHA_MINIMA	FECHA_MAXIMA	Contador_Frases	Contador_Frases_Unicas	Contador_Frases_Nulas
Notas_Evolucion	2019-01-02 00:00:00.000	2020-08-30 00:00:00.000	306780	173711	0
Notas_Ingreso	2019-01-06 19:22:14.000	2020-08-30 12:17:04.000	1440	486	0
TOTAL	2019-01-02 00:00:00.000	2020-08-30 12:17:04.000	308220	174197	0

Se cargaron un total de 308.220 registros, de las cuales 174.197 son frases únicas (esto indica que varias frases con igual estructura son usadas en diferentes pacientes).

Tabla 17 - Cantidad de registros por paciente

Contador_Pacientes	Total_Registros	Registros_Minimos_x_Paciente	Registros_Promedio_x_Paciente	Registros_Maximos_x_Paciente	Fecha_Minima_Registros	Fecha_Maxima_x_Registros
980	308220	6	314	4252	2019-01-02 17:37:49.000	2020-08-30 23:48:22.000

Se cargó información de 980 pacientes, en promedio un paciente tiene 314 registros y hasta un máximo de 4245 registros. (Ver tabla 18)



	min	median	std	mean	max	skew	kurtosis
Nro_registros	5.0	168.0	486.573047	313.040816	4245.0	5.271591	35.665761

Ilustración 8 - Histograma y Distribución del # de registros por Paciente

El histograma y el sesgo revela que tenemos una distribución asimétrica positiva de la cantidad de registros por paciente y la curtosis de tipo Leptocúrtica indica que hay concentración de registros, la moda se posiciona en 59. (Ver ilustración 8). En resumen, son más los pacientes que tienen pocas notas que los pacientes que tienen muchas notas.

Tabla 18 - Cantidad de registros promedio por PDE

Contador_PDE	Total_Registros	Registros_Minimos_x_PDE	Registros_Promedio_x_PDE	Registros_Maximos_x_PDE	Fecha_Minima_Registros	Fecha_Maxima_x_Registros
314	306780	0	973	5814	2019-01-02 17:37:49.000	2020-08-30 23:48:22.000

Se cargó información de 314 enfermeros que crearon al menos una NN. En promedio un PDE registra 973 notas, sin embargo, se encontraron PDE con hasta 5.814 notas. (Ver tabla 19).

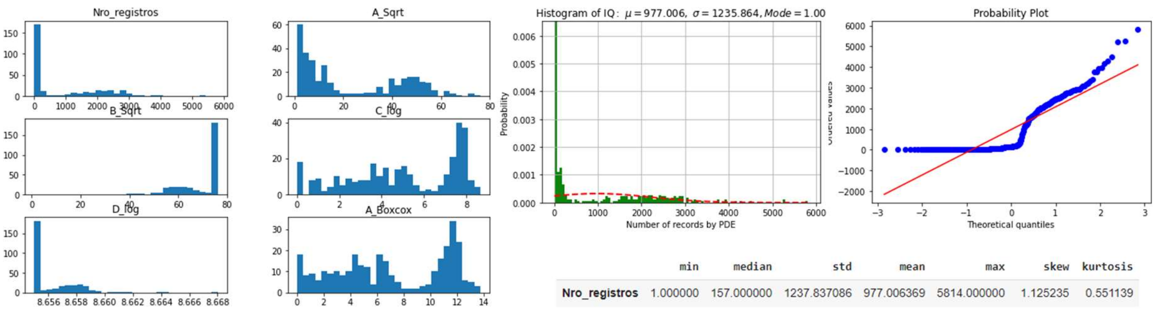


Ilustración 9 - Histograma y Distribución del # de registros por PDE

El histograma y el sesgo revela que tenemos una distribución asimétrica positiva de la cantidad de registros por PDE y la curtosis de tipo Leptocúrtica indica que hay concentración de registros, la moda se posiciona en 1. (Ver ilustración 9). En resumen, son pocos los PDE que registran NN (es posible que se designe a un pequeño grupo de personas para registrar las notas), el resto de los enfermeros registran las notas individualmente.

Tabla 19 - Porcentaje de errores ortográficos (palabras únicas)

Tipo_Palabra	Total_Palabras	%_Participacion
Palabra_Correcta	31436	0.2916006
Palabra_Incorrecta	76369	0.7083994
TOTAL	107805	1

Se encontró que el 70% de las palabras únicas usadas en la muestra cargada tienen algún tipo problema ortográfico, de siglas, acrónimos y abreviaciones. (Ver tabla 20)

Tabla 20 - Similitud de palabras

Palabras	Contador_Palabras	Frecuencia_Palabras	%_Participacion_Palabras_Similares	%_Participacion_Frecuencia_Palabras_Similares
CON PALABRAS SIMILARES	45735	31728622	0.4146644	0.6056409
SIN PALABRAS SIMILARES	64559	20659881	0.5853356	0.394359
TOTAL	110294	52388503	1	1

Resumen del análisis preliminar de los datos

- Sobre el total de los registros por paciente, se encontró un 7% de frases duplicadas.
- Se encontró un 60% de problemas ortográficos.
- Las palabras únicas equivalen al 55% del total de palabras, el 45% restante son palabras con similitud con otra, usualmente mal escritas (Ver tabla 21), son palabras diferentes en su escritura, pero en esencia son la misma. Este tipo de problemas son frecuentes hasta un 77%. Se entiende mejor en la ilustración 10 y 11. La figura 10 muestra que dos palabras similares (iguales, pero mal escritas) puede estar presente 2.500 veces, tres palabras mal escritas 1500 veces y así sucesivamente. La figura 11 ilustra el impacto en frecuencia que puede tener, por ejemplo, una palabra que mal escrita tiene 48 formas de escribirse, y la sumatoria en frecuencia es de 4.500.000. Con lo anterior se evidencia que el mayor problema que se tiene en los datos es la falta de ortografía y de estándar en los acrónimos, sinónimos y abreviaciones, lo que dificulta análisis posteriores.
- Ej. La palabra "Paciente", tiene 177 posibles formas de escribirse.

Tabla 21 - Cuantificando los errores de escritura

GroupByToken_Key	Token_Value	Token_Value_Pre_Fixed	Token_Value_Fixed	Token_Value_Fixed_Type	Token_DataLenght	Frequency	Participation	Active
74043	paciente	paciente	Paciente	2	NULL	822139	0.01569311	1
70816	pacinete	pacinete	paciente	4	NULL	1764	3.36715E-05	1
30733	apciente	apciente	paciente	4	NULL	646	1.233095E-05	1
30591	apaciente	apaciente	paciente	2	NULL	480	9.162312E-06	1
72301	pacient	pacient	paciente	4	NULL	455	8.685109E-06	1
104844	paciete	paciete	paciente	4	NULL	423	8.074288E-06	1
72263	pacienet	pacienet	paciente	4	NULL	394	7.520732E-06	1
74388	paceinte	paceinte	paciente	4	NULL	332	6.337266E-06	1
74417	pacietne	pacietne	paciente	4	NULL	263	5.020184E-06	1
76422	pciente	pciente	paciente	4	NULL	249	4.75295E-06	1

- Participación de palabras similares (Cantidad y Frecuencia) sobre el total de los datos

Ilustración 10 - Participación de palabras con similitud sobre la cantidad total de palabras

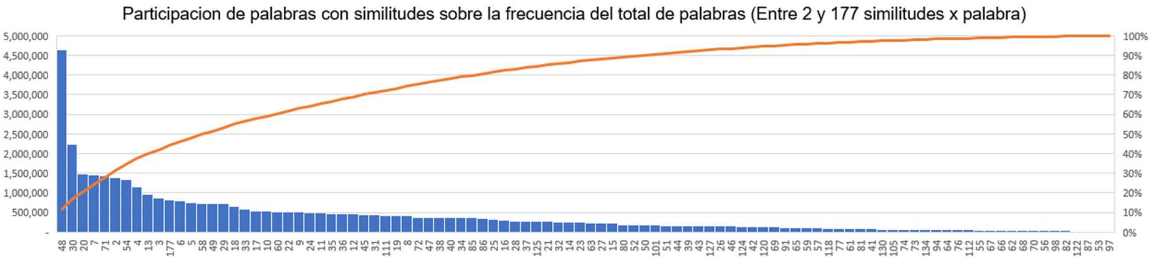


Ilustración 11 - Participación de palabras con similitud sobre la frecuencia total de palabras

Rondas de Limpieza

La metodología implementada descompuso las frases en tokens y estos fueron limpiados en dos rondas de limpieza, cada una con tres etapas descritas en la tabla 6. Este proceso es indispensable para la cuantificación de las notas. La ilustración 12 evidencia que la cantidad de palabras únicas paso de 110.294 a 44.794 (59% de reducción), esto debido a la corrección ortográfica, estandarización de siglas, acrónimos y abreviaciones, remoción de stop words y lematización. La frecuencia pasó de 52.4 MM a 34 MM (35% de disminución). De no aplicarse esta metodología, el tiempo de procesamiento y calidad de los resultados es pobre.

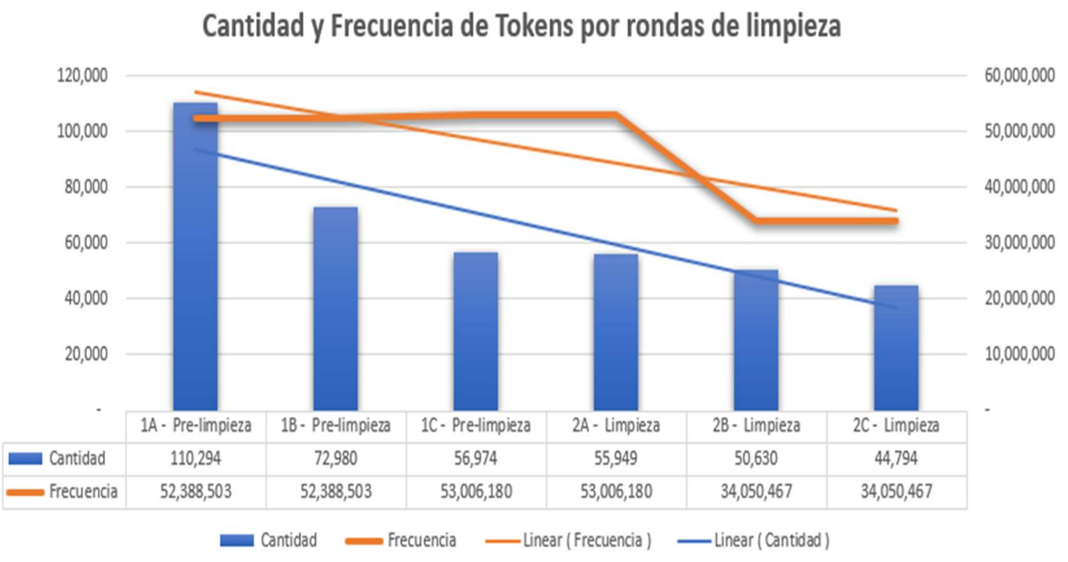


Ilustración 12 - Cantidad y Frecuencia de tokens por rondas de limpieza

Construcción de Categorías

En esta etapa del proceso, los datos cuentan con una limpieza del 90%, comparada con los datos originales. Para la construcción de categorías se usó dos enfoques, construcción de categorías con el experto (PDE) y mediante modelamiento de tópicos (TM).

Construcción de la categoría con el experto

Según los criterios de los PDE en reuniones realizadas previamente, se diseñó una tabla para almacenar categorías de contenido de las notas mediante palabras clave (metadata), con el fin de hacer más fácil la identificación de categorías y subcategorías tratadas en las NN, como se aprecia en la tabla 22. Por otro lado, para complementar la metadata, se usó Topic Modeling mediante el algoritmo de LDA, y se identificaron agrupación de temas alrededor de mismos tópicos, los cuales no eran evidentes para el PDE. (Ver ilustración 13, 14 y 15). Con estos dos enfoques, se alimentó la tabla 22 y se ve materializada en la ilustración 16.

Tabla 22 - Categorías identificadas por los PDE

TAG	PRIMERA PALABRA	SEGUNDA PALABRA	SINONIMO
DISPOSITIVOS MEDICOS	SONDA	VESICAL	
		NASOGASTRICA	
		OROGASTRICA	
		GASTROSTOMIA	
	CATETER	PERIFERICO	ACCESO VENOSO PERIFERICO
		PICC	
		CENTRAL	
	LINEA	ARTERIAL	
	TUBO	OROTRAQUEAL	
CIRUGIA	PROCEDIMIENTO	QUIRURGICO	
PROCEDIMIENTO	HEMOCULTIVO		
	CATETERISMO	VESICAL	
	CURACION	HERIDAS	
		CATETER	
SOPORTE NUTRICIONAL	SOPORTE	METABOLICO	
	NUTRICION	ENTERAL	
	NUTRICION	PARENTERAL	
MOMENTOS DE CUIDADO	INGRESO DEL	PACIENTE	
	VALORACION DE	ENFERMERIA	PLAN DE ATENCION DE ENFERMERIA
	PLANEACION	NECESIDADES	BASICAS
	RECIBO	TURNO	
	ENTREGA DE	TURNO	
	ADMINISTRACION DE	MEDICAMENTOS	
	MOMENTO	EGRESO	
VALORACION DE PIEL	**ESCALA DE**	**BRADEN**	
	RIESGO	ALTO	
	RIESGO	MODERADO	
	RIESGO	BAJO	
VALORACION DE RIESGO DE CAIDA	**ESCALA**	**MACDEMS**	**ESCALA DE RIESGO DE CAIDA PEDIATRICA**
	RIESGO	ALTO	
	RIESGO	MEDIO	
	RIESGO	BAJO	
	ESCALA DE	**MORSE**	**ESCALA DE RIESGO DE CAIDA ADOLESCENTES Y ADULTOS**
	RIESGO	ALTO	
	RIESGO	MEDIO	
	RIESGO	BAJO	
GRADO DE DEPENDENCIA	ESCALA DE	BARTHEL	O INDICE DE BARTHEL
CONTROL DE SIGNOS VITALES	TENSION	ARTERIAL	TA / TAM
	TEMPERATURA		
	FRECUENCIA	CARDIACA	FC
	DOLOR		
	FRECUENCIA	RESPIRATORIA	FR
	SATURACION	OXIGENO	SAT O2

Tabla 23 - Tabla en base de datos con las categorías definidas por el PDE y TP

KeyWord_Key	KeyWord_TAG	KeyWord_TAG_Priority	KeyWord_Category	KeyWord_Operator1	KeyWord1	KeyWord_Operator2	KeyWord2	KeyWord_Operator3	KeyWord3	KeyWord_Operator4	KeyWord4	KeyWord_Operator5	KeyWord5	KeyWord_Priority	KeyWord_Score	KeyWord_Active
77	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MACDEMS	LIKE	RIESGO CAIDA	LIKE	MEDIO BAJO	NULL	NULL	NULL	NULL	3	NULL	1
78	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MORSE	LIKE	RIESGO CAIDA	LIKE	MEDIO BAJO	NULL	NULL	NULL	NULL	3	NULL	1
95	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MACDEMS	LIKE	RIESGO CAIDA	LIKE	MEDIO ALTO	NULL	NULL	NULL	NULL	5	NULL	1
96	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MORSE	LIKE	RIESGO CAIDA	LIKE	MEDIO ALTO	NULL	NULL	NULL	NULL	5	NULL	1
94	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	RIESGO	LIKE	CAIDA	LIKE	MEDIO ALTO	NULL	NULL	NULL	NULL	5	NULL	1
85	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MACDEMS	LIKE	RIESGO CAIDA	LIKE	MEDIO	NULL	NULL	NULL	NULL	4	NULL	1
86	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MORSE	LIKE	RIESGO CAIDA	LIKE	MEDIO	NULL	NULL	NULL	NULL	4	NULL	1
88	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MACDEMS	LIKE	RIESGO CAIDA	LIKE	MODERADO	NULL	NULL	NULL	NULL	4	NULL	1
89	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MORSE	LIKE	RIESGO CAIDA	LIKE	MODERADO	NULL	NULL	NULL	NULL	4	NULL	1
100	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MACDEMS	LIKE	RIESGO CAIDA	LIKE	ALTO	NULL	NULL	NULL	NULL	6	NULL	1
101	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MORSE	LIKE	RIESGO CAIDA	LIKE	ALTO	NULL	NULL	NULL	NULL	6	NULL	1
76	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	RIESGO	LIKE	CAIDA	LIKE	MEDIO BAJO	NULL	NULL	NULL	NULL	3	NULL	1
70	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MACDEMS	LIKE	RIESGO CAIDA	LIKE	BAJO	NULL	NULL	NULL	NULL	2	NULL	1
71	VALORACION DE RIESGO DE CAIDA	30	NULL	LIKE	MORSE	LIKE	RIESGO CAIDA	LIKE	BAJO	NULL	NULL	NULL	NULL	2	NULL	1

Construcción de la categoría con un proceso no supervisado de NLP

Para la construcción de estas categorías, se usó la técnica de modelamiento de tópicos (TM) usando el algoritmo de LDA. En este un proceso no NLP no supervisado, donde fueron analizadas las notas limpias y lematizadas con la librería *gensim* de Python. Con ella se procesaron las palabras en unigramas, bigramas y trigramas, se calcularon el número óptimo de TM por n gramas (Ver ilustración 13) para decidir cuál se comportaba mejor para complementar las categorías que ya se tenían con el usuario experto.

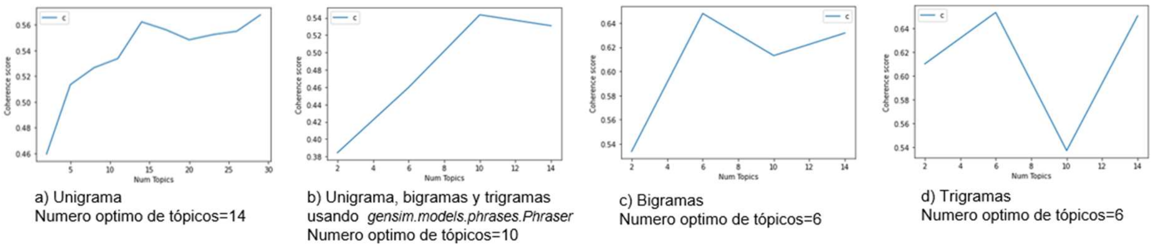


Ilustración 13 - Número óptimo de tópicos para unigramas, bigramas y trigramas

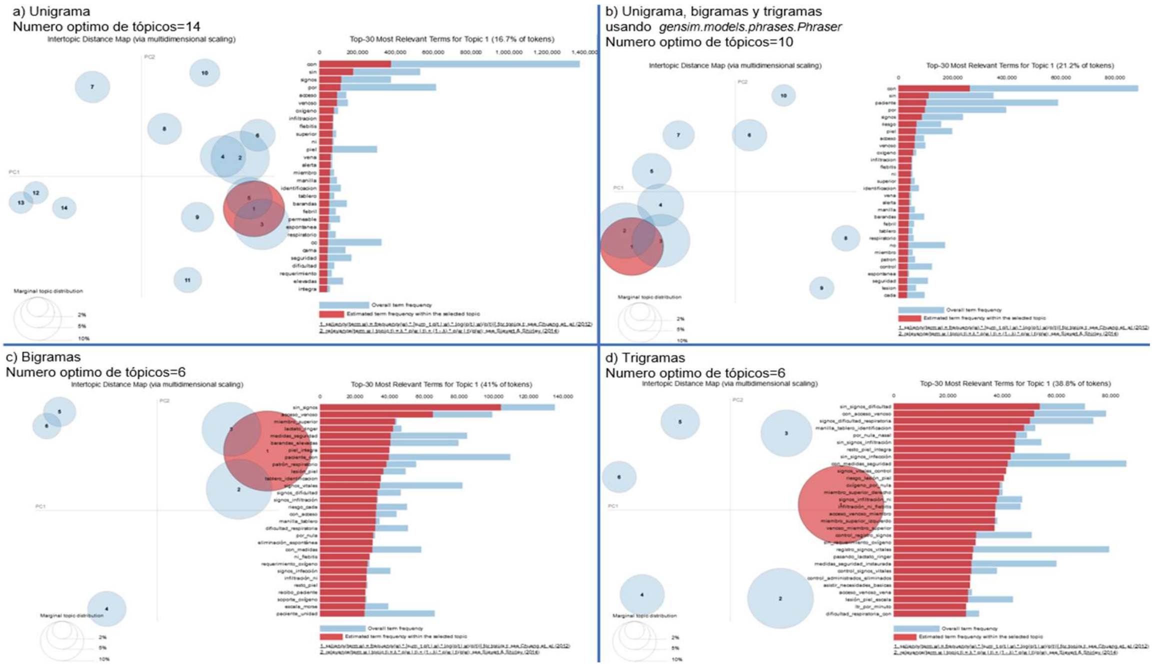


Ilustración 14 - Distribución de tópicos por tipo de n grama

Ilustración 15 - n-gramas más frecuentes por tópico

Dominancia por tópico

Las tablas 24, 25 y 26 representan los tópicos más representativos por nota, para la tabla 24, el tópico dominante por unigramas es el número 10 (Columna Dominant Topic), para la tabla 25, el tópico dominante por Unigramas, Bigramas y Trigramas es el número 6, para la tabla 26, el tópico dominante por bigramas es el número 5.

Tabla 24 - Topic Modeling para Unigramas

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	3	0.7271	sin, con, enfermera, unidad, realiza, ronda, s...	[paciente, quien, continua, unidad, cuidados, ...
1	1	0	0.8174	con, cc, por, sonda, matter, sin, piel, unidad...	[control, glucometria, que, reporta, mg, dl, a...
2	2	0	0.4657	con, cc, por, sonda, matter, sin, piel, unidad...	[recibo, paciente, anos, da, con, siguientes, ...
3	3	12	0.9226	cc, cut, equipo, insumos, con, matter, isopani...	[solicita, farmacia, equipo, xl, buretrol, nav...
4	4	4	0.3433	realiza, jefe, turno, signos, toma, vitales, p...	[realiza, ronda, enfermera, paciente, unidad, ...
5	5	3	0.6362	sin, con, enfermera, unidad, realiza, ronda, s...	[realiza, ronda, enfermera, paciente, unidad, ...
6	6	8	0.6538	con, enfermera, piel, por, sin, manejo, signos...	[uci, jorge, hernan, verano, bolivar, nota, en...
7	7	0	0.3020	con, cc, por, sonda, matter, sin, piel, unidad...	[realiza, ronda, enfermera, paciente, unidad, ...
8	8	10	0.8839	mg, va, gr, iv, medicamentos, administracion, ...	[registro, administra, medicamentos, previa, v...
9	9	10	0.5112	mg, va, gr, iv, medicamentos, administracion, ...	[administra, hidromorfona, mg, intravenosa, pa...
10	10	10	0.9071	mg, va, gr, iv, medicamentos, administracion, ...	[administra, gluconato, calcio, ampolla, intra...
11	11	8	0.2965	con, enfermera, piel, por, sin, manejo, signos...	[paciente, extubacion, sin, complicaciones, re...
12	12	0	0.4658	con, cc, por, sonda, matter, sin, piel, unidad...	[realiza, ronda, enfermera, paciente, unidad, ...
13	13	0	0.6571	con, cc, por, sonda, matter, sin, piel, unidad...	[recibo, paciente, unidad, cuidados, intensivo...
14	14	6	0.8542	con, realiza, manos, lavado, procedimiento, es...	[nota, realiza, matter, central, yugular, izqu...
15	15	10	0.7512	mg, va, gr, iv, medicamentos, administracion, ...	[nota, administracion, medicamentos, previa, m...

Tabla 25 - Topic Modeling para Unigramas, Bigramas y Trigramas

	NN_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	7	0.4897	paciente, sin, realiza, con, enfermera, signos...	[paciente, quien, continua, unidad, cuidados, ...
1	1	6	0.8218	con, cc, por, paciente, signos, derecho, sonda...	[control, glucometria, que, reporta, mg, dl, a...
2	2	6	0.7536	con, cc, por, paciente, signos, derecho, sonda...	[recibo, paciente, anos, da, con, siguientes, ...
3	3	2	0.9250	con, cut, realiza, matter, equipo, manos, por,...	[solicita, farmacia, equipo, xl, buretrol, nav...
4	4	0	0.7809	paciente, realiza, con, cambio, piel, por, pos...	[realiza, ronda, enfermera, paciente, unidad, ...
5	5	0	0.5483	paciente, realiza, con, cambio, piel, por, pos...	[realiza, ronda, enfermera, paciente, unidad, ...
6	6	9	0.6492	con, paciente, enfermera, por, del, no, piel, ...	[uci, jorge, hernan, verano, bolivar, nota, en...
7	7	0	0.5618	paciente, realiza, con, cambio, piel, por, pos...	[realiza, ronda, enfermera, paciente, unidad, ...
8	8	5	0.8875	folio, abre, registro, medicamentos, verificac...	[registro, administra, medicamentos, previa, v...
9	9	4	0.5289	va, mg, oral, gr, elementos, complicaciones, b...	[administra, hidromorfona, mg, intravenosa, pa...
10	10	5	0.4685	folio, abre, registro, medicamentos, verificac...	[administra, gluconato, calcio, ampolla, intra...
11	11	0	0.4882	paciente, realiza, con, cambio, piel, por, pos...	[paciente, extubacion, sin, complicaciones, re...
12	12	6	0.6492	con, cc, por, paciente, signos, derecho, sonda...	[realiza, ronda, enfermera, paciente, unidad, ...
13	13	6	0.8804	con, cc, por, paciente, signos, derecho, sonda...	[recibo, paciente, unidad, cuidados, intensivo...
14	14	2	0.9281	con, cut, realiza, matter, equipo, manos, por,...	[nota, realiza, matter, central, yugular, izqu...
15	15	5	0.9250	folio, abre, registro, medicamentos, verificac...	[nota, administracion, medicamentos, previa, m...

Tabla 26 - Topic Modeling para Bigramas

	NN_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	1	0.9100	ronda_enfermera, signos_vitales, jefe_turno, s...	[paciente_quien, quien_continua, continua_unid...
1	1	2	0.8515	sin_signos, acceso_venoso, miembro_superior, l...	[control_glucometria, glucometria_que, que_rep...
2	2	2	0.9105	sin_signos, acceso_venoso, miembro_superior, l...	[recibo_paciente, paciente_41, 41_años, años_d...
3	3	1	0.7587	ronda_enfermera, signos_vitales, jefe_turno, s...	[solicita_farmacia, farmacia_equipo, equipo_50...
4	4	1	0.8920	ronda_enfermera, signos_vitales, jefe_turno, s...	[realiza_ronda, ronda_enfermera, enfermera_pac...
5	5	1	0.9636	ronda_enfermera, signos_vitales, jefe_turno, s...	[realiza_ronda, ronda_enfermera, enfermera_pac...
6	6	3	0.8918	paciente_con, diagnóstico_enfermera, sin_signo...	[uci_16, 16_jorge, jorge_hernan, hernan_verano...
7	7	1	0.6762	ronda_enfermera, signos_vitales, jefe_turno, s...	[realiza_ronda, ronda_enfermera, enfermera_pac...
8	8	0	0.8612	administración_medicamentos, abre_folio, momen...	[registro_administra, administra_medicamentos,...
9	9	5	0.3963	va_oral, mg_va, bajo_técnica, proteccion_perso...	[administra_hidromorfona, hidromorfona_0.4, 0....
10	10	0	0.9163	administración_medicamentos, abre_folio, momen...	[18_00, 00_administra, administra_[illegible]uconato, g...
11	11	3	0.4980	paciente_con, diagnóstico_enfermera, sin_signo...	[paciente_extubación, extubación_sin, sin_comp...
12	12	2	0.7750	sin_signos, acceso_venoso, miembro_superior, l...	[realiza_ronda, ronda_enfermera, enfermera_pac...
13	13	2	0.9879	sin_signos, acceso_venoso, miembro_superior, l...	[recibo_paciente, paciente_unidad, unidad_cuid...
14	14	5	0.6468	va_oral, mg_va, bajo_técnica, proteccion_perso...	[nota_realiza, realiza_matter, matter_central,...
15	15	5	0.5869	va_oral, mg_va, bajo_técnica, proteccion_perso...	[nota_administración, administración_medicamen...
16	16	5	0.7220	va_oral, mg_va, bajo_técnica, proteccion_perso...	[dexametasona_mg, mg_iv]
17	17	5	0.7916	va_oral, mg_va, bajo_técnica, proteccion_perso...	[hidromorfona_0.4, 0.4_mg, mg_iv]
18	18	3	0.9015	paciente_con, diagnóstico_enfermera, sin_signo...	[nota_enfermera, enfermera_turno, turno_noche,...
19	19	0	0.9357	administración_medicamentos, abre_folio, momen...	[abre_folio, folio_registro, registro_administ...

Notas de enfermería más representativas por tópico

Las tablas 27, 28 y 29 muestran las 10 primeras notas de enfermería en donde las palabras usadas están fuertemente correlacionadas a los tópicos detectados por el algoritmo LDA, la tabla 27 está compuesta por unigramas y cada registro es el que mejor representa cada tópico, lo mismo para la tabla 28 (unigramas, bigramas y trigramas) y la tabla 29 (bigramas).

Tabla 27 - Documentos más representativos por tópico para unigramas

	Topic_Num	Topic_Perc_Contrib	Keywords	Text
0	0	0.9916	con, cc, por, sonda, matter, sin, piel, unidad...	[queda, paciente, unidad, cuidados, intensivos...
1	1	0.9140	con, por, trauma, lesion, abdominal, pop, frac...	[paciente, masculino, anos, edad, con, medicos...
2	2	0.9643	va, mg, oral, no, complicaciones, gr, acetamin...	[omeprazol, mg, vo, posaron, mg, vo, levotirox...
3	3	0.9814	sin, con, enfermera, unidad, realiza, ronda, s...	[pasa, ronda, enfermera, paciente, continua, b...
4	4	0.9807	realiza, jefe, turno, signos, toma, vitales, p...	[jefe, turno, con, previa, tecnica, toma, labo...
5	5	0.9643	riesgo, del, cada, dolor, escala, no, valoraci...	[modelo, funcional, enfermera, momento, cuidad...
6	6	0.9867	con, realiza, manos, lavado, procedimiento, es...	[matter, venoso, central, realiza, matter, cen...
7	7	0.9810	control, signos, piel, vitales, vigilar, enfer...	[horas, cateterismos, musicales, cada, edema, ...
8	8	0.9814	con, enfermera, piel, por, sin, manejo, signos...	[am, valoracion, fisica, sin, signos, dificult...
9	9	0.9929	con, sin, signos, por, acceso, venoso, oxigeno...	[ofrecio, dieta, va, oral, ordenada, queda, pa...

Tabla 28 - Documentos más representativos por tópico para Unigramas, Bigramas y Trigramas

	Topic_Num	Topic_Perc_Contrib	Keywords	Text
0	0	0.9823	paciente, realiza, con, cambio, piel, por, pos...	[realiza, cambio, posicion, ms, lubricacion, c...
1	1	0.9743	cada, riesgo, escala, valoracion, con, refiere...	[momento, cuidado, administracion, medicamento...
2	2	0.9935	con, cut, realiza, matter, equipo, manos, por,...	[realiza, matter, central, subclavia, izquierd...
3	3	0.9938	con, sin, paciente, por, signos, riesgo, piel,...	[recibo, paciente, anos, da, con, siguientes, ...
4	4	0.9769	va, mg, oral, gr, elementos, complicaciones, b...	[carvedilol, tableta, mg, dosis, mg, va, oral,...
5	5	0.9526	folio, abre, registro, medicamentos, verificac...	[dra, borrego, realiza, rasurado, zona, quirur...
6	6	0.9933	con, cc, por, paciente, signos, derecho, sonda...	[recibo, paciente, unidad, con, medidas, segur...
7	7	0.9833	paciente, sin, realiza, con, enfermera, signos...	[paciente, descansa, intervalos, largos, con, ...
8	8	0.9795	medicamentos, administracion, paciente, admini...	[momento, cuidado, modelo, funcional, enfermer...
9	9	0.9800	con, paciente, enfermera, por, del, no, piel, ...	[linfocitos, hemoglobina, hematocrito, plaquet...

Tabla 29 - Documentos más representativos por tópico para Bigramas

	Topic_Num	Topic_Perc_Contrib	Keywords	Text
0	0	0.9861	administración_medicamentos, abre_folio, momen...	[dr_reyes, reyes_inicia, inicia_anestesia, ane...
1	1	0.9929	ronda_enfermera, signos_vitales, jefe_turno, s...	[realiza_ronda, ronda_enfermera, enfermera_pac...
2	2	0.9969	sin_signos, acceso_venoso, miembro_superior, l...	[recibo_paciente, paciente_86, 86_años, años_1...
3	3	0.9975	paciente_con, diagnóstico_enfermera, sin_signo...	[horas_con, con_un, un_volumen, volumen_250, 2...
4	4	0.9860	administración_medicamentos, brinda_educación,...	[inicia_unidades, unidades_plasma, plasma_as, ...
5	5	0.9891	va_oral, mg_va, bajo_técnica, proteccion_perso...	[doripenem_500, 500_mg, mg_ssn, ssn_0.9, 0.9_1...

Notas de enfermería más representativas por tópico

Como resultado del TM, se usaron los unigramas, bigramas y trigramas más comunes por tópico para complementar la tabla por categorías (Ver tabla 23). Los mejores tópicos se encuentran al usar bigramas y trigramas (Ver ilustraciones 13a, 13b, 14a, 14b). Este paso revela nueva información no conocida por los enfermeros y es una pieza clave para la creación de metadatos. Las categorías resultantes son empleadas para alimentar el proceso de categorización de las notas y son usadas en la visualización de los datos.

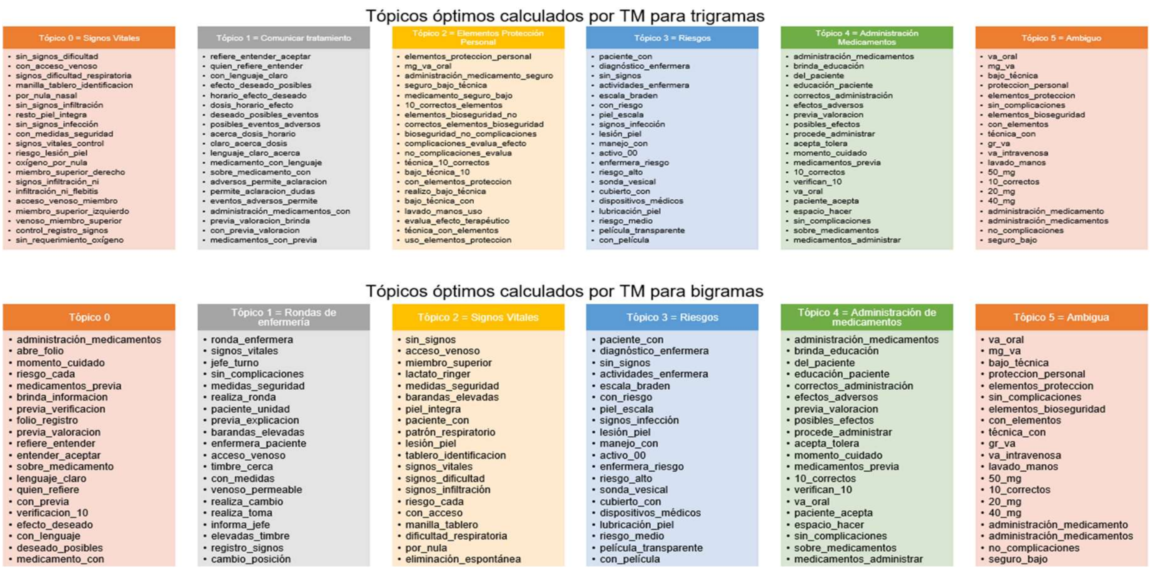


Ilustración 16 - Construcción de categorías a partir de los tópicos encontrados en los trigramas y bigramas

Construcción del puntaje de las notas de enfermería

Con las nuevas interacciones entre enfermero-paciente detectadas y las NN lematizadas, se cuantificó el valor de ellas usando el algoritmo BM25. Usándolo, se cuantificó el valor de cada nota sobre el total de las notas. Luego, se sumó el valor de cada palabra dentro del texto para construir un score por nota, al final, se ponderó el valor de cada nota sobre la nota que tiene el mayor puntaje (Ver tabla 30 resaltada en azul) para obtener el porcentaje individual.

Adicional al score por nota, se creó un segundo, pero esto solo uso las categorías y tópicos extraídos de las notas de la sección anterior *"Se cuantificó solo el contenido importante"*. Para ello se creó una cuantificación a partir de Gold Standards Corpus (GSC). Básicamente con el puntaje del BM25 que tiene cada palabra dentro de una nota, se totalizó el valor de cada una, pero solo sumando las palabras clave que hacen parte de las categorías preseleccionadas por los PDE y TM, es decir, solo se cuantifican las palabras que están marcadas como relevantes. También se ponderó el valor de cada registro con la nota que tiene el mayor GSC. (Ver tabla 30 resaltada en verde).

Tabla 30 - Puntaje generado por NN

Token_SRC_Key	ScoreNote	Score_GoldStandard	Count_GoldStandard	ScoreNote_Over_Best_Score	Score_GoldStandard_Over_Best_Score
169453	4924.34	229.18	38	1.00	0.96
251373	3910.05	182.39	29	0.79	0.76
104254	3443.87	169.22	34	0.70	0.71
70157	3388.95	238.86	36	0.69	1.00
291443	3284.02	155.68	24	0.67	0.65

Token_SRC_Key	ScoreNote	Score_GoldStandard	Count_GoldStandard	ScoreNote_Over_Best_Score	Score_GoldStandard_Over_Best_Score
70157	3388.95	238.86	36	0.69	1.00
169453	4924.34	229.18	38	1.00	0.96
264400	2738.81	218.34	52	0.56	0.91
169455	2234.98	218.22	29	0.45	0.91
212469	2119.61	208.56	29	0.43	0.87

Visualización del puntaje

Una vez se tiene el puntaje Okapi BM25 y Gold Standard (GS) del paso anterior, se procede a visualizar cada descripción mediante la herramienta **Power BI**, para ello se diseñó un tablero que permite proyectar la nota original (17b) y la limpia (17c), además se pintan dos nubes de palabras, una con los (unigramas, bigramas y trigramas) más representativos (ilustración 17e) y la otra donde se resaltan los tokens con más fuerza dentro del texto (17f). También se muestra la metadata de la nota (17g) y se imprime dinámicamente el BM25 y GS (17d) calculado para todo registro. Las anotaciones se pueden buscar fácilmente usando el calendario dinámico a mano izquierda del panel (17a).

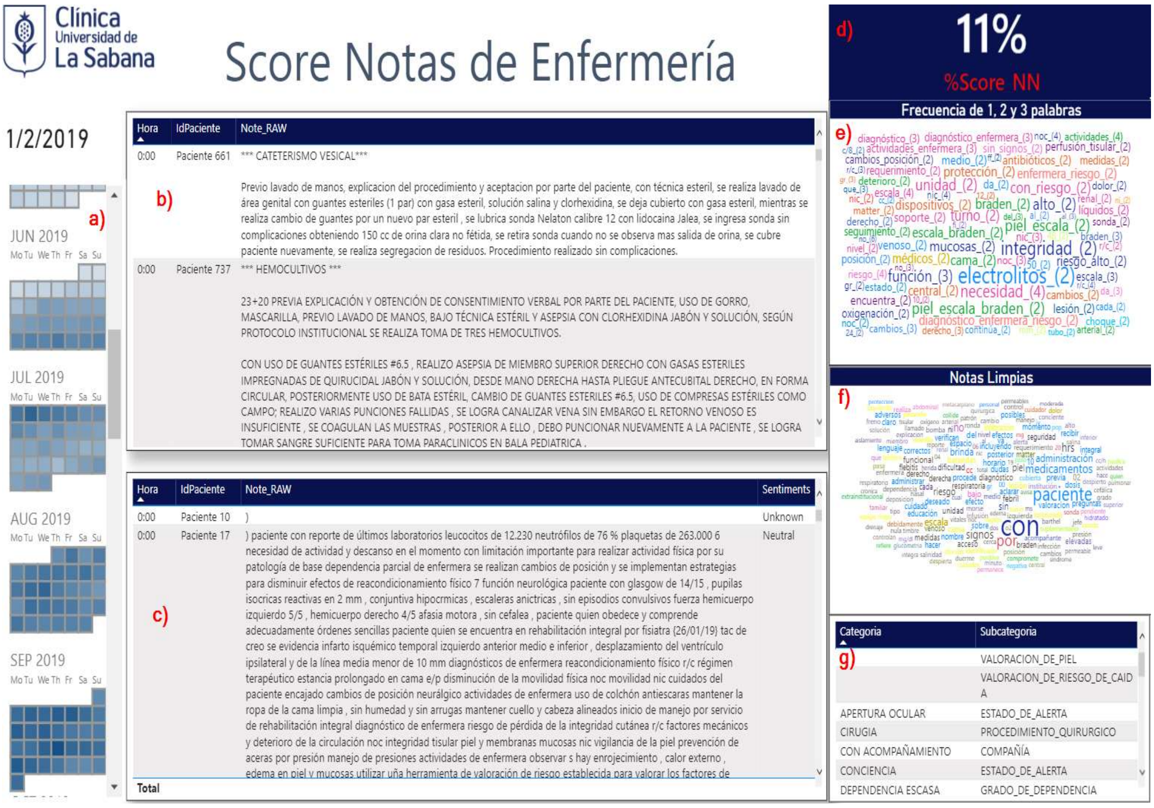


Ilustración 17 - Score Notas de Enfermería. a) Calendario dinámico, b) Nota original, c) Nota arreglada, d) Score por nota / Score con Gold Standard, e) Nube de palabras por unigramas/ bigramas/trigramas, f) Nube de palabras por token, g) Metadata

Análisis de perfiles de los Profesionales de Enfermería

Desde el cargue de datos se conoce el código que diligenció la NN, con este se logra individualizar todo registro y mediante otros cálculos posibles por la analítica de texto y por algoritmos usados en NLP es posible elaborar el perfilamiento de autor (AP). Con el análisis textual se consigue medir (la cantidad y longitud de palabras) mínimas, máximas y promedio que registra cada enfermero. Con estas se pueden estimar medidas centrales como desviación estándar, moda, media, mediana, curtosis y asimetría. Similar al ejemplo previo, también se permite el cómputo para las notas y frases usadas. Otra información que se puede obtener en la etapa de AP es la exploración paratextual de los textos (uso de símbolos, mayúsculas, minúsculas, números y caracteres especiales), la diversidad léxica, el número de pacientes atendidos, porcentaje de errores de ortografía (sin tener en cuenta acentos) y los puntajes por nota y por GSC obtenidos del paso anterior. Con todas las métricas obtenidas, es factible agrupar los PDE según su rendimiento. La segmentación habilita la detección de grupo de enfermeros que requieren mayor entrenamiento en la redacción de las NN y el construir indicadores de mejora continua. (Encontrar todas las variables usadas en la tabla 31).

Tabla 31 - Variables para el perfilamiento del profesional de enfermería

ITEM	DESCRIPCION
TOTAL_Interactions_By_PDE	Total interacciones enfermero-paciente
MIN_Token_DataLenght_RAW	Longitud mínima palabras usadas por el enfermero
AVG_Token_DataLenght_RAW	Longitud promedio palabras usadas por el enfermero
MAX_Token_DataLenght_RAW	Longitud máxima palabras usadas por el enfermero
MODE_Token_DataLenght_RAW	Moda de la longitud de las palabras usadas por el enfermero
MEDIAN_Token_DataLenght_RAW	Mediana de la longitud de las palabras usadas por el enfermero
STDEV_Token_DataLenght_RAW	Desviación estándar de la longitud de las palabras usadas por el enfermero
VAR_Token_DataLenght_RAW	Varianza de la longitud de las palabras usadas por el enfermero
SKEWNESS_Token_DataLenght_RAW	Asimetría de la longitud de las palabras usadas por el enfermero
KURTOSIS_Token_DataLenght_RAW	Curtosis de la longitud de las palabras usadas por el enfermero
TOTAL_Token_Uppercase_RAW	Total de palabras en mayúscula digitadas por el enfermero
Percent_Uppercase	Porcentaje de palabras en mayúscula usadas por el enfermero sobre el total de sus palabras digitadas
TOTAL_Token_HasSpecialCharacter_RAW	Total de palabras con caracteres especiales digitadas por el enfermero
Percent_SpecialCharacters	Porcentaje de palabras con caracteres especiales digitadas por el enfermero sobre el total de sus palabras digitadas
TOTAL_Token_IsNumeric_RAW	Total de palabras con números digitadas por el enfermero
Percent_Numbers	Porcentaje de palabras con números digitadas por el enfermero sobre el total de sus palabras digitadas
TOTAL_Token_Value_Misspelling_RAW	Total de palabras con algún error ortográfico digitado por el enfermero
Percent_Misspelling	Porcentaje de palabras con algún error ortográfico digitado por el enfermero sobre el total de sus palabras digitadas
TOTAL_Words	Total de palabras digitadas por el enfermero
TOTAL_Unique_Words	Total de palabras únicas digitadas por el enfermero
TOTAL_Lexical_Diversity	Porcentaje de palabras únicas sobre total de palabras digitadas por el enfermero
TOTAL_Notes_By_PDE	Total de notas diligenciadas por el enfermero
MIN_Words	Palabras mínimas usadas por el enfermero
AVG_Words	Palabras promedio usadas por el enfermero
MAX_Words	Palabras máximas usadas por el enfermero

Desempeño de los profesionales de enfermería

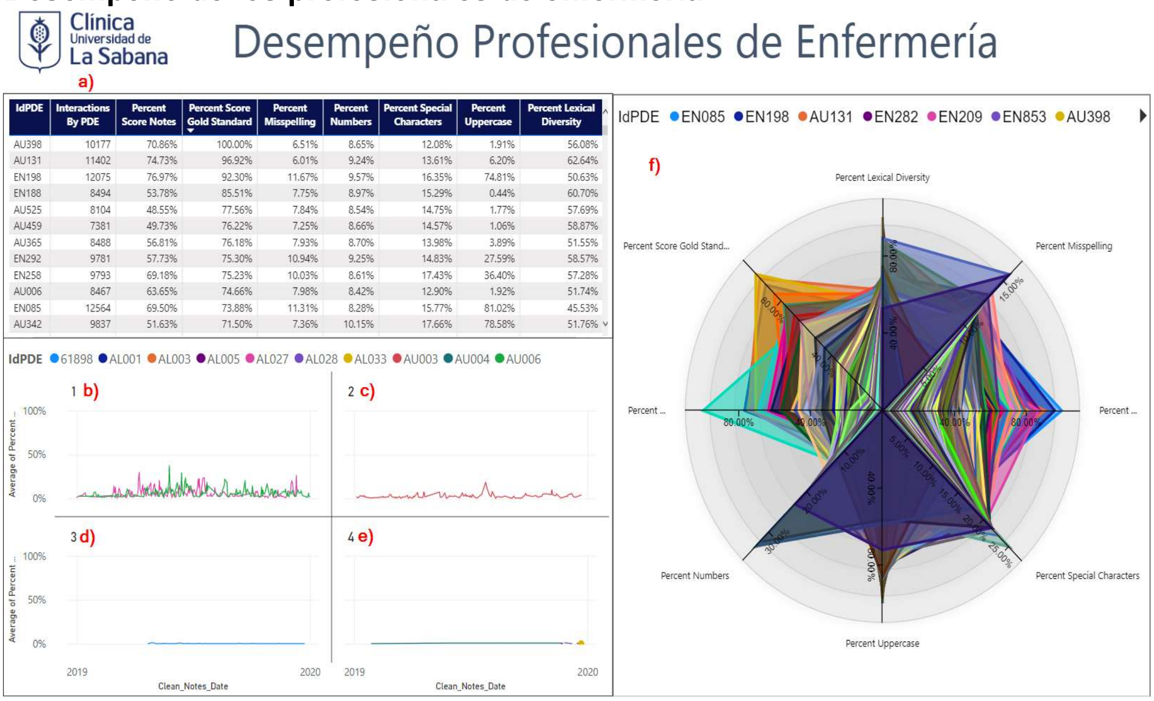
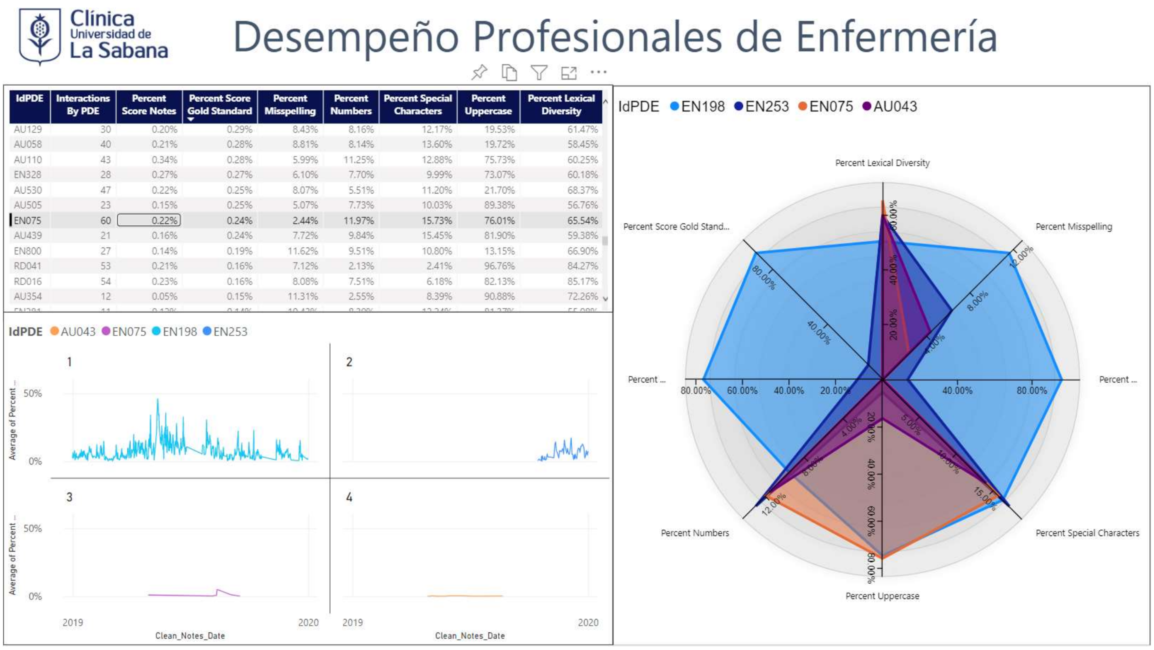


Ilustración 18 - Desempeño de los profesionales de enfermería. a) Cuadro resumen de los enfermeros, b) primer cuadrante de interacciones, c) segundo cuadrante de interacciones, d) tercer cuadrante de interacciones, e) Cuarto cuadrante de interacciones, f) Gráfico de radar del performance del PDE

La ilustración 18 es una ayuda visual realizada en **Power BI** que permite comparar el performance de los enfermeros, para ello se cuenta con una tabla resumen (Ilustración 18a) donde se evalúan métricas como Porcentaje de puntaje de notas (Calculado sumando el total acumulado del puntaje de las notas escritas por el enfermero y comparado contra el enfermero con puntaje más alto), porcentaje con el Gold Estándar - GS (Calculado sumando el total acumulado del puntaje de las GS de las notas escritas por el enfermero y comparado contra el enfermero con puntaje más alto), Porcentaje de errores de ortografía (ponderación del total de errores ortográficos sobre el total de las palabras usadas), Porcentaje de números usados (ponderación del total de números usados sobre el total de palabras usadas), Porcentaje de caracteres especiales (ponderación del total de caracteres especiales usados (Ver tabla 32) sobre el total de palabras usadas), porcentaje de mayúsculas (ponderación del total de palabras mayúsculas usadas sobre el total de palabras usadas) y porcentaje de diversidad léxica (Se calcula con el número de palabras únicas usadas vs el total de palabras usadas por el enfermero). El primer cuartil (18b), el segundo cuartil (18c), el tercer cuartil (18d) y el cuarto cuartil (18f) son 4 secciones del gráfico donde se dividen los enfermeros, en función de la cantidad de notas registradas. En el primer segmento se muestran los PDE que tienen mayor número de interacciones hasta llegar a la sección 4, que muestra el que tiene pocas interacciones. La visualización permite comparar el performance de múltiples enfermeros a la vez, lo cual es útil para segmentar enfermeros que requieren más entrenamiento para la redacción de las NN. (Ver ilustración 19).



IdPDE	Interactions By PDE	Percent Score Notes	Percent Score Gold Standard	Percent Misspelling	Percent Numbers	Percent Special Characters	Percent Uppercase	Percent Lexical Diversity
AU129	30	0.20%	0.29%	8.43%	8.16%	12.17%	19.53%	61.47%
AU058	40	0.21%	0.28%	8.81%	8.14%	13.60%	19.72%	58.45%
AU110	43	0.34%	0.28%	5.99%	11.25%	12.88%	75.73%	60.25%
EN328	28	0.27%	0.27%	6.10%	7.70%	9.99%	73.07%	60.18%
AU530	47	0.22%	0.25%	8.07%	5.51%	11.20%	21.70%	68.37%
AU505	23	0.15%	0.25%	5.07%	7.73%	10.03%	89.38%	56.76%
EN075	60	0.22%	0.24%	2.44%	11.97%	15.73%	76.01%	65.54%
AU439	21	0.16%	0.24%	7.72%	9.84%	15.45%	81.90%	59.38%
EN800	27	0.14%	0.19%	11.62%	9.51%	10.80%	13.15%	66.90%
RD041	53	0.21%	0.16%	7.12%	2.13%	2.41%	96.76%	84.27%
RD016	54	0.23%	0.16%	8.08%	7.51%	6.18%	82.13%	85.17%
AU354	12	0.05%	0.15%	11.31%	2.55%	8.39%	90.88%	72.26%

Ilustración 19 - Comparación entre enfermeros

CONCLUSIONES

En esta tesis se ha demostrado la posibilidad de generar nuevo conocimiento a partir de las notas de enfermería en español de la Clínica Universidad de La Sabana, mediante técnicas de procesamiento de datos y analítica de texto. El nuevo conocimiento proviene del hecho que la cantidad de texto que generan los profesionales de enfermería excede la capacidad de análisis de cualquier ser humano. Así, estas técnicas han permitido la construcción automática de indicadores que permiten la mejora continua de los protocolos de enfermería y del contenido que deberían tener las notas de enfermería.

Es innegable el aporte que tiene la analítica de texto para construir metainformación que puede ser usada para el mejoramiento continuo de los procesos y de las personas que participan. Así, con la metodología propuesta fue posible extraer, transformar, y cargar los registros de las notas de enfermería en texto saludable para la construcción de indicadores de gestión del cuidado.

Los primeros desafíos encontrados fue conseguir las herramientas y algoritmos adecuadas que funcionaran en lengua castellana. Esto debido a la falta de recursos lingüísticos en español, en especial para temas especializados, como la medicina y la enfermería. En la metodología también fue esencial comprender las características de las buenas notas de enfermería y; las historias de usuario jugaron un papel importante para ello las historias de usuario condujeron a la identificación de los problemas existentes, para así diseñar las reglas, el porcentaje de saneamiento y orden necesario para procesar las anotaciones para que tengan sentido.

El proceso más costoso en todo el proyecto fue la limpieza de datos. Este paso requirió del 80% del total del tiempo. Sin embargo, son indispensables para encontrar información oculta y generar nueva información, por ejemplo, las interacciones reales entre enfermero-paciente. El Topic modeling fue un factor determinante para distinguir la agrupación de las palabras y de qué manera estas forman tópicos, necesarias para la creación de categorías.

El texto limpio y lematizado habilito la cuantificación de las notas de enfermería, con ella fue factible cuantificar el valor de una nota en función de los tokens utilizados. Adicionalmente, una segunda medida fue generada, de acuerdo con categorías y claves presentes en los registros (contenido deseable). Otro subproducto del preprocesamiento y escorización de las notas fue el haber posible realizar el análisis de perfil de los profesionales de enfermería, La cantidad de palabras escritas, el porcentaje de errores, el score que tienen las observaciones y KeyWords empleadas habilitaron el poder calcular el performance del enfermero. Con esta información, es plausible crear planes de entrenamiento para mejorar las habilidades de redacción de los profesionales de enfermería que lo requieren.

Lo anterior hay que disponibilizarlo y la visualización de los datos es el medio ideal para mostrar estas nuevas interacciones, conocer el score de las notas y poder comparar el performance entre los PDE, de una manera intuitiva, amigable y sencilla. En resumen, Este tipo de soluciones permite entender la situación actual de cómo se están diligenciando las notas de enfermería, abstrayendo del texto métricas cuantificables para la toma de decisiones.

Es técnicamente viable llevar a la práctica esta solución en la CUS, se requiere de base de datos para persistir la data, lenguaje Python para la ejecución de ciertos algoritmos especializados, Herramientas de visualización para desplegar la nueva información generada, usar los recursos lingüísticos para comparar, corregir, y estandarizar las palabras con problemas, ETLs para la automatización de los procesos y por supuesto, apoyar al usuario experto para validar la efectividad de las correcciones realizadas.

Por último, esta investigación genero un marco de trabajo, desde la transformación de los datos hasta su uso por parte de los usuarios finales y sugirió las herramientas, arquitecturas tecnológicas, pasos y buenas prácticas para lograrlo.

TRABAJO FUTURO

El actual trabajo sienta las bases para futuros estudios, la cuantificación de las NN puede ser usadas para implementar otros algoritmos como redes neuronales profundas, con estas, es posible entrenar modelos de ML para que aprendan lo que debe contener una buena NN, así mismo, este tipo de soluciones se pueden integrar en tiempo real con otras aplicaciones hospitalarias, para permitir al PDE contar con herramientas que le permitan realizar una mejor descripción y redacción en cuanto el estatus y novedades del paciente, y que esta pueda ser usada para futuros análisis o auditorías.

Otro campo de estudio es la posibilidad de crear un macroproyecto para la construcción de recursos lingüísticos en idioma español enfocado a la medicina. La CUS y la Universidad de la Sabana pueden ser pioneros en este espacio, las futuras aplicaciones se pueden dar a nivel universitario y clínico en el territorio colombiano con opción de ser usados internacionalmente.

Otra información no incluida y con potencial uso son los datos y archivos no estructurados como historias clínicas, radiográficas, resultados clínicos, entre otros. Pueden ser usados para correlacionar patrones que permitan al personal médico tomar mejores decisiones.

REFERENCIAS BIBLIOGRÁFICAS

1. Hyun S, Cooper C. Application of Text Mining to Nursing Texts: Exploratory Topic Analysis. Compute Inform Nurse. 2020 Oct;38(10):475-482. doi: 10.1097/CIN.0000000000000681. PMID: 33044316.
2. Bjarnadottir, R. I., & Lucero, R. J. (2018). What Can We Learn about Fall Risk Factors from EHR Nursing Notes? A Text Mining Study. Egems (generating Evidence & Methods to Improve Patient Outcomes), 6(1), 21. DOI: http://doi.org/10.5334/egems.237
3. Martos López, Daniel. (2021). Detección de eventos adversos en historiales clínicos mediante Procesamiento del Lenguaje Natural Master Thesis, Universidad Nacional de Educación a Distancia (España). Escuela Técnica Superior de Ingeniería Informática. Departamento de Inteligencia Artificial
4. Tellería, Carlos & Ilarri, Sergio & Sánchez, Carlos. (2020). Text Mining of Medical Documents in Spanish: Semantic Annotation and Detection of Recommendations. 197-208. 10.5220/0010059101970208.
5. Galatzan, Benjamin J. PhD, RN; Carrington, Jane M. PhD, RN, FAAN; Gephart, Sheila PhD, RN, FAAN Testing the Use of Natural Language Processing Software and Content Analysis to Analyze Nursing Hand-off Text Data, CIN: Computers, Informatics, Nursing: March 27, 2021 - Volume Publish Ahead of Print - Issue - doi: 10.1097/CIN.0000000000000732
6. Chang HM, Huang EW, Hou IC, Liu HY, Li FS, Chiou SF. Using a Text Mining Approach to Explore the Recording Quality of a Nursing Record System. J Nurs Res. 2019 Jun;27(3):e27. doi: 10.1097/jnr.0000000000000295. PMID: 30694223; PMCID: PMC6553963.
7. Ministerio de Educación. Resolución número 1995 de 1999. Disponible en https://www.minsalud.gov.co/Normatividad_Nuevo/RESOLUCI%C3%93N%201995%20DE%201999.pdf Consultado: 16 de junio, 2021
8. Ministerio de Educación. Ley 911 de 2004. (octubre 5) Diario Oficial. No. 45.693 de 6 de octubre de 2004. Disponible en: http://www.mineducacion.gov.co/1621/articles-105034_archivo_pdf.pdf Consultado: 16 de junio, 2021.
9. Lamprea Reyes, L., Bejarano López, A., & Nieto González, M. (2016). "Adaptación del paciente y su cuidador familiar durante el cambio de turno de enfermería en el servicio de hospitalización de la Clínica Universidad de La Sabana". Universidad De La Sabana. Retrieved from https://intellectum.unisabana.edu.co/handle/10818/26040
10. Moreno Sandoval, Luis Gabriel & Beltrán-Herrera, Paola & Vargas, Jaime & Sánchez-Barriga, Carolina & Pomares Quimbaya, Alexandra & Alvarado, Jorge & Garcia, Juan. (2017). CSL: A Combined Spanish Lexicon - Resource for Polarity Classification and Sentiment Analysis. 288-295. 10.5220/0006336402880295.
11. https://www.elsevier.com/es-es/connect/medicina/escala-de-coma-de-glasgow
12. Salinas, Martha & Moreno Sandoval, Luis Gabriel & Pomares Quimbaya, Alexandra. (2018). Philosophy of Technology in Affective Computing and Social Network Analysis. 10.18687/LACCEI2018.1.1.35.

13. Moreno Sandoval, Luis Gabriel & Beltrán-Herrera, Paola & Vargas, Jaime & Sánchez-Barriga, Carolina & Pomares Quimbaya, Alexandra & Alvarado, Jorge & Garcia, Juan. (2017). CSL: A Combined Spanish Lexicon - Resource for Polarity Classification and Sentiment Analysis. 288-295. 10.5220/0006336402880295.
14. Moreno Sandoval, Luis Gabriel & Mendoza, Joan & Puertas, Edwin & Duque-Marín, Arturo & Pomares Quimbaya, Alexandra & Alvarado, Jorge. (2018). Age Classification from Spanish Tweets - The Variable Age Analyzed by using Linear Classifiers. 10.5220/0006811102750281.
15. Moreno Sandoval, Luis Gabriel & Puertas, Edwin & Plaza-Del-Arco, F.M. & Pomares Quimbaya, Alexandra & Alvarado, Jorge & López, L.. (2019). Celebrity Profiling on Twitter using Sociolinguistic Features Notebook for PAN at CLEF 2019.
16. Puertas, Edwin & Moreno Sandoval, Luis Gabriel & Plaza-Del-Arco, F.M. & Alvarado, Jorge & Pomares Quimbaya, Alexandra & López, L.. (2019). Bots and Gender Profiling on Twitter using Sociolinguistic Features Notebook for PAN at CLEF 2019.
17. Puertas, Edwin & Moreno Sandoval, Luis Gabriel & Redondo, · & Alvarado, Jorge & Pomares Quimbaya, Alexandra. (2021). Detection of Sociolinguistic Features in Digital Social Networks for the Detection of Communities. Cognitive Computation. 13. 20. 10.1007/s12559-021-09818-9.
18. Moreno Sandoval, Luis Gabriel & Pomares Quimbaya, Alexandra & Alvarado, Jorge. (2021). Celebrity profiling through linguistic analysis of digital social networks. Computational Social Networks. 8. 10.1186/s40649-021-00097-w.
19. Oliveira, Nicollas & Pisa, Pedro & Andreoni, Martin & Medeiros, Dianne & Menezes, Diogo. (2021). Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. Information. 12. 38. 10.3390/info12010038.
20. MUNANA-RODRIGUEZ, J. E. y RAMIREZ-ELIAS, A.. Escala de coma de Glasgow: origen, análisis y uso apropiado. Enferm. univ [online]. 2014, vol.11, n.1, pp.24-35. ISSN 2395-8421.
21. 21. Lamprea Reyes, L., Bejarano López, A., & Nieto González, M. (2016). "Adaptación del paciente y su cuidador familiar durante el cambio de turno de enfermería en el servicio de hospitalización de la Clínica Universidad de La Sabana". Universidad De La Sabana. Retrieved from https://intellectum.unisabana.edu.co/handle/10818/26040.
22. Matlab (Ed.). (n.d.). ¿Qué es un n-grama? ¿Qué es un n-grama? - MATLAB. Retrieved May 25, 2022, from https://la.mathworks.com/discovery/ngram.html
23. Introducción al análisis de texto. (2020). Retrieved 26 May 2022, from https://old.tacosdedatos.com/analisis-texto
24. Ordoñez, Hugo. (2011). STEMMING EN ESPAÑOL PARA DOCUMENTOS RECUPERADOS DE LA WEB* STEMMING IN THE SPANISH LANGUAGE FOR DOCUMENTS RECOVERED FROM THE WEB. Revista Unimar. 58.

25. IBM Docs. (2022). Retrieved 26 May 2022, from https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=techniques-co-occurrence-rules
26. Introducción al topic modeling con Gensim (I): fundamentos y preprocesamiento de textos. (2021). Retrieved 30 May 2022, from https://elmundodelosdatos.com/topic-modeling-gensim-fundamentos-preprocesamiento-textos/
27. Blei, David & Ng, Andrew & Jordan, Michael & Lafferty, John. (2003). Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03 Latent Dirichlet Allocation.
28. Tf-idf - Wikipedia, la enciclopedia libre. (2022). Retrieved 30 May 2022, from https://es.wikipedia.org/wiki/Tf-idf#:~:text=Tf%2Didf%20(del%20ingl%C3%A9s%20Term,un%20documento%20en%20una%20colecci%C3%B3n.
29. Vilares, Jesús. (2008). El modelo probabilístico: características y modelos derivados. Revista General de Información y documentación. ISSN 1132-1873. 18. 345-363.
30. Wissler, Lars & Almashraee, Mohammed & Monett, Dagmar & Paschke, Adrian. (2014). The Gold Standard in Corpus Annotation. 10.13140/2.1.4316.3523.
31. Métodos de prospectiva, MICMAC : La prospective. (2022). Retrieved 6 July 2022, from http://es.laprospective.fr/Metodos-de-prospectiva/Los-programas/67-Micmac.html#:~:text=El%20m%C3%A9todo%20Micmac%20para%20multiplicaci%C3%B3n,del%20material%20de%20prospectiva%20estrat%C3%A9gica.
32. erwin Data Modeler. Industry-Leading Data Modeling Tool | erwin, Inc. (2022). Retrieved 6 July 2022, from https://www.erwin.com/products/erwin-data-modeler/?gclid=Cj0KCQjw5ZSWBhCVARIsALERCvy4E46rOwzgAndrZ3daYAYZ0Aj_RfV8jh5QxfjoT9BrE5ADp-AOskwaAu-kEALw_wcB&gclsrc=aw.ds
33. Rueda, J. (2019). CRISP-DM: una metodología para minería de datos en salud healthdataminer.com. Retrieved 31 July 2022, from https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

ANEXOS

Tabla 32 - Caracteres para remover según ASCII / ISO 8859-1

Dec	Value	Description	TREATMENT
33	!	Exclamation mark	AVOID
34	"	Quotation mark/Double quote	AVOID
39	'	Apostrophe/Single quote	AVOID
63	?	Question mark	AVOID
92	\	Reverse solidus/Backslash	AVOID
94	^	Circumflex accent/Caret	AVOID
96	`	Grave accent	AVOID
126	~	Tilde	AVOID
128			AVOID
129			AVOID
130			AVOID
131			AVOID
132			AVOID
133			AVOID
134			AVOID
135			AVOID
136			AVOID
137			AVOID
138			AVOID
139			AVOID
140			AVOID
141			AVOID
142			AVOID
143			AVOID
144			AVOID
145			AVOID
146			AVOID
147			AVOID
148			AVOID
149			AVOID
150			AVOID
151			AVOID
152			AVOID

Dec	Value	Description	TREATMENT
153			AVOID
154			AVOID
155			AVOID
156			AVOID
157			AVOID
158			AVOID
159			AVOID
160		No-break space	AVOID
161	¡	Inverted exclamation mark	AVOID
162	¢	Cent sign	AVOID
163	£	Pound sign	AVOID
164	¤	Currency sign	AVOID
165	¥	Yen/Yuan sign	AVOID
166	¦	Broken bar	AVOID
167	§	Section sign	AVOID
168	¨	Diaeresis	AVOID
169	©	Copyright sign	AVOID
170	ª	Feminine ordinal indicator	AVOID
171	«	Left-pointing double angle quotation mark	AVOID
172	¬	Not sign	AVOID
173	-	Soft hyphen	AVOID
174	®	Registered trademark sign	AVOID
175	¯	Macron	AVOID
177	±	Plus-minus sign	AVOID
178	²	Superscript two	AVOID
179	³	Superscript three	AVOID
180	´	Acute accent	AVOID
181	µ	Micro sign (mu)	AVOID
182	¶	Pilcrow sign	AVOID
183	·	Middle dot	AVOID
184	¸	Cedilla	AVOID
185	¹	Superscript one	AVOID
186	º	Masculine ordinal indicator	AVOID

Dec	Value	Description	TREATMENT
187	»	Right-pointing double angle quotation mark	AVOID
188	¼	Vulgar fraction one quarter	AVOID
189	½	Vulgar fraction one half	AVOID
190	¾	Vulgar fraction three quarters	AVOID
191	¿	Inverted question mark	AVOID
192	À	Latin capital letter A with grave	AVOID
193	Á	Latin capital letter A with acute	AVOID
194	Â	Latin capital letter A with circumflex	AVOID
195	Ã	Latin capital letter A with tilde	AVOID
196	Ä	Latin capital letter A with diaeresis	AVOID
197	Å	Latin capital letter A with ring above	AVOID
198	Æ	Latin capital letter AE	AVOID
199	Ç	Latin capital letter C with cedilla	AVOID
200	È	Latin capital letter E with grave	AVOID
201	É	Latin capital letter E with acute	AVOID
202	Ê	Latin capital letter E with circumflex	AVOID
203	Ë	Latin capital letter E with diaeresis	AVOID
204	Ì	Latin capital letter I with grave	AVOID
205	Í	Latin capital letter I with acute	AVOID
206	Î	Latin capital letter I with circumflex	AVOID
207	Ï	Latin capital letter I with diaeresis	AVOID
208	Ð	Latin capital letter ETH	AVOID
210	Ò	Latin capital letter O with grave	AVOID
211	Ó	Latin capital letter O with acute	AVOID
212	Ô	Latin capital letter O with circumflex	AVOID
213	Õ	Latin capital letter O with tilde	AVOID
214	Ö	Latin capital letter O with diaeresis	AVOID
215	×	Multiplication sign	AVOID
216	Ø	Latin capital letter O with stroke	AVOID
217	Ù	Latin capital letter U with grave	AVOID
218	Ú	Latin capital letter U with acute	AVOID
219	Û	Latin capital letter U with circumflex	AVOID
220	Ü	Latin capital letter U with diaeresis	AVOID

Dec	Value	Description	TREATMENT
221	Ý	Latin capital letter Y with acute	AVOID
222	Þ	Latin capital letter THORN	AVOID
223	ß	Latin small letter sharp s	AVOID
224	à	Latin small letter a with grave	AVOID
225	á	Latin small letter a with acute	AVOID
226	â	Latin small letter a with circumflex	AVOID
227	ã	Latin small letter a with tilde	AVOID
228	ä	Latin small letter a with diaeresis	AVOID
229	å	Latin small letter a with ring above	AVOID
230	æ	Latin small letter ae	AVOID
231	ç	Latin small letter c with cedilla	AVOID
232	è	Latin small letter e with grave	AVOID
233	é	Latin small letter e with acute	AVOID
234	ê	Latin small letter e with circumflex	AVOID
235	ë	Latin small letter e with diaeresis	AVOID
236	ì	Latin small letter i with grave	AVOID
237	í	Latin small letter i with acute	AVOID
238	î	Latin small letter i with circumflex	AVOID
239	ï	Latin small letter i with diaeresis	AVOID
240	ð	Latin small letter eth	AVOID
242	ò	Latin small letter o with grave	AVOID
243	ó	Latin small letter o with acute	AVOID
244	ô	Latin small letter o with circumflex	AVOID
245	õ	Latin small letter o with tilde	AVOID
246	ö	Latin small letter o with diaeresis	AVOID
247	÷	Division sign/Obelus	AVOID
248	ø	Latin small letter o with stroke	AVOID
249	ù	Latin small letter u with grave	AVOID
250	ú	Latin small letter u with acute	AVOID
251	û	Latin small letter u with circumflex	AVOID
252	ü	Latin small letter u with diaeresis	AVOID
253	ý	Latin small letter y with acute	AVOID
254	þ	Latin small letter thorn	AVOID
255	ÿ	Latin small letter y with diaeresis	AVOID

Tabla 33 - Frases repetidas y metadatos encontrados

Fecha Evolucion	Hora Evolucion	Nota Evolucion
2019-01-20	07:00:12	05:00 PACIENTE QUIEN CONTINUA EN UNIDAD DE CUIDADOS INTENSIVOS CON MEDIDAS DE SEGURIDAD INSTAURADOS POR INSTITUCION, BARANDAS ELEVADAS HEMODINAMICAMENTE ESTABLE, AFEBRIL, BAJO MEDIDAS DE SEDOANALGESIA, SIN SOPORTE VASOPRESOR, SE REALIZA BAÑO GENERAL EN CAMA, LUBRICACION DE PIEL, CAMBIO DE TENDIDOS, SIN COMPLICACIONES 06:00 CONTROL DE GLUCOMETRIA QUE REPORTA 73 MG/DL SE AVISA A JEFE DE TURNO POR LO CUAL SE MIDE RESIDUO GASTRICO DE 20 CC POR LO CUAL SE SUBE NUTRICION ENTERAL A LA META NUTRICIONAL 50 CC/HORA, QUEDA PACIENTE EN UNIDAD DE CUIDADOS INTENSIVOS CON MEDIDAS DE SEGURIDAD INSTAURADOS POR INSTITUCION, BARANDAS ELEVADAS, MANILLA Y TABLERO DE IDENTIFICACION, CON ALTO RIESGO DE LESION DE PIEL SEGÚN ESCALA DE BRADEN Y RIESGO MEDIO DE CAIDA SEGÚN ESCALA DE MORSE, A LA VALORACION CEFALOCAUDAL PACIENTE BAJO MEDIDAS DE SEDOANALGESIA, SIN SOPORTE VASOPRESOR, TUBO OROTRAQUEAL CONECTADA A VENTILACION MECANICA POR PARAMETROS ESTABLECIDOS POR TERAPIA RESPIRATORIA, ACOPLADO AL VENTILADOR, SONDA OROGASTRICAPASANDO PERATIVE 50 CC/HORA, CATETER CENTRAL YUGULAR IZQUIERDO CUBIERTO CON PELICULA DE FIJACION LIMPIO SIN SIGNOS DE INFECCION PASANDO FENTANYL 12 CC/HORA, MIDAZOLAM 13.3 CC/HORA, LACTATO DE RINGER 100 CC/HORA, REPOSICION DE POTASIO 4 CC/HORA, ACCESO VENOSO EN MIEMBRO SUPERIOR DERECHO CUBIERTO CON APOSITO DE FIJACION LIMPIO SIN SIGNOS DE FLEBITIS, LINEA ARTERIAL RADIAL DERECHA CUBIERTA CON PELICULA DE FIJACION LIMPIA SIN SIGNOS DE INFECCION, FUNCIONAL SIN SIGNOS DE HIPOPERFUSION, VENDAJE LAMINADO EN MIENBRO SUPERIOR IZQUIERDO CON ADECUADO LLENADO CAPILAR Y PERFUCION DISTAL,HERIDA QUIRURGICA LUMBAR CUBIERTA CON GASA Y FIXOMULL, TUBO A TORAX DERECHO CUBIERTO CON GASA Y FIXOMULL CONECTADO A PLEUROVACK QUE DURANTE LA NOCHE DRENO 150 CC, SONDA VESICAL CONECTADA A CYSTOFLO, FLICTENAS EN TALONES DERECHO E IZQUIERDO, CONTINUA EN MANEJO MEDICO.
2019-08-20	16:00:00	13+00 Paciente continua con indicacion de nada via oral , en espera de realizacion de EVDA Y COLONOSCOPIA 14+30 se asisten necesidades basicas de la paciente , continua en compañia de familiar bajo medidas de seguridad instauradas por la institucion . 15+30 previa explicacion y aceptacion por parte de paciente y familiar , se realiza toma y control de signos vitales y se registran en el sistema , los cuales se encuentran dentro de parámetros normales se informa a la jefe de turno , paciente con escala de dolor 3/10 16+00 se pasa ronda de enfermeria paciente continua en habitacion ,Sin signos de dificultad respiratoria, con orden de nada via oral , con un primer acceso venoso periférico en vena basilica superior izquierda , para paso de lactato de ringer a 120cc/h, y un segundo acceso venoso en vena basilica superior derecha conectado a catéter salinizado para el paso de medicamentos , accesos limpios,

Registros duplicados

Metadatos
El paciente tiene 2 visitas, a las 5 y 6 am. Solo hay una nota.

Tabla 34 - Referencias

N°	TOPIC	AUTHORS	METHODOLGY	SUMMARY	DOI	Key Words
1	Applicaton of Text Ming to Nursing Texts	Sookyung Hyun, Cheryl Cooper	Exploratory Analysis	La investigación uso técnicas de minería de texto sobre récords electrónicos de salud (EHR) y los clasifico en 40 tópicos	10.1097/CIN.0000000000000681	TM, EHR, Topic
2	Testing the Use of Natural Language Processing Software and Content Analysis to Analyze Nursing Hand-off Text Data	Benjamin J. Galatzan, Jane M. Camington, Sheila Gephart	Exploratory Analysis	Este estudio usa técnicas de Machine Learning como LIWC and Renz para analizar la transferencia de conocimiento entre enfermeros	10.1097/CIN.0000000000000732	ML, LIWC
3	Using a Text Mining approach to explore the recording quality of a nurse record system	Hsiu-Mei CHANG, Ean-Weng HUANG, I-Ching HOU, Hsiu-Yun LIU, Fang-Shan LI, Shwu-Fen CHIOU	Exploratory Analysis	Uso de herramientas somo SAS y minería de texto para validar la calidad de las notas de enfermería y hacer sugerencias de como mejorarlas	10.1097/jnr.0000000000000295	TM, NN Quality
4	What Can We Learn about Fall Risk Factors from EHR Nursing Notes? A text Mining Study	Ragnhildur I, Bjarnadottir and Robert J. Lucero	Exploratory Analysis	Este estudio hace uso de técnicas de machine learning como N-gram y Lexycom para entender el sentido de las palabras en las notas de enfermería	10.5334/egems.237.s1	ML, N-grams, Lexycom
5	A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data	Caitlin Dreisbach, Theresa A. Koleck, Philip E. Bourne and Suzanne Bakken	A comprehensive literature	Uso del procesamiento del lenguaje natural (NLP) y la minería de textos en su aplicación a la extracción y procesamiento de síntomas en el texto electrónico escrito por el paciente (ePAT).	10.1016/j.ijmedinf.2019.02.008	NLP, TM, ePAT
6	Using nursing notes to improve clinical outcome prediction in intensive care patients: A retrospective cohort study	Kexin Huang, Tamryn F. Gray, Santiago Romero-Brufau, James A. Tulsky, Charlotta Lindvall	Research and Applications	Este estudió analiza las notas de enfermería y médicos de pacientes de UCI para pronosticar con ellas al deceso o no del paciente.	10.1093/jamia/ocab051	Human Analysis
7	The influence of integrated electronic medical records and computerized nursing notes on nurses' time spent in documentation	Tracy Yee, Jack Needleman, Marjorie Pearson, Patricia Parkerton, Melissa Parkerton, Joelle Wolstein	Statisctics Analysis	Hace una a análisis estadístico del tiempo que las enfermeras gastan en diligenciar las notas de enfermería con y sin uso de notas electrónicas	10.1097/NXN.0b013e31824af835	Statistics Analysis
8	Applying artificial intelligence technology to support decisionmaking in nursing: A case study in Taiwan	Pei-Hung Liao, Pei-Ti Hsu, William Chu, Woei-Chyn Chu	Article	Uso de Inteligencia Artificial con redes neuronales y otras herramientas estadísticas como SPSS para analizar los textos de enfermería	10.1177/1460458213509806	AI, Neuronal Networks
9	Challenges and opportunities beyond structured data in analysis of electronic health records	Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, Fred Godtliebsen	Advanced Review	Analiza el potencial uso de los datos estructurados y no estructurados de los registros hospitalarios para atención de los pacientes, teniendo en cuenta los retos para implementarlo	10.1002/wics.1549	Unstructured Data
10	Theoretical Considerations of Ethics in Text Mining of Nursing Documents	Hanna Suominen, Tuija Lehtikunnas, Barbro Back, Helena Karsten, Tapio Salakoski, Sanna Salanterä	Article	Consideraciones éticas al usar minería de texto con las notas de enfermería		Statistics Analysis